Characteristics of Datasets and Models Reuse Patterns on the Hugging Face Platform

Park, Jaihyun Indiana University Indianapolis, USA | jaipark@iu.edu

Yoon, Ayoung Indiana University Indianapolis, USA | ayyoon@iu.edu

ABSTRACT (150 WORDS)

This study empirically explores how Natural Language Processing (NLP) and Computer Vision (CV) datasets and models are reused in the Hugging Face community. We find that NLP tasks - such as Zero-shot-classification, Sentence-similarity, and Feature-extraction - require more diverse datasets compared to CV tasks on average. On the other hand, NLP datasets were reused less frequently than CV datasets. In addition, CV models were reused frequently to develop other models compared to NLP models. In conclusion, NLP models reused diverse datasets for training, while CV datasets and models were reused more and layered up together to develop other models. This study contributes to the understudied area of dataset and model reuse in computing and the broader data reuse subfield under Information Science.

KEYWORDS

Data reuse, Data sharing, Data lifecycle, Artificial intelligence

INTRODUCTION

Artificial Intelligence (AI) has surprised people by answering questions with human-like language and knowledge (e.g., ChatGPT) and by generating artistic visuals in domains traditionally perceived to be a field of human creativity (e.g., DALL-E 3 (Huang & Goh, 2024)). These advancements in AI technologies are in part attributed to the model and data sharing practices. Researchers in Machine Learning (ML) competed for the higher ground using benchmark datasets (Park & Jeoung, 2022) to demonstrate that their models perform better. Models and datasets created through competitions are made public to demonstrate their technical competencies (Zhou et al., 2018) and eventually contributed to the subsequent emerging fields that entail applied research using social media data (Park et al., 2024) and historical data (Park & Cordell, 2023a; Park & Cordell 2025; Griebel et al., 2024). However, while existing literature has focused on data sharing and reuse practices in academia, the reuse of datasets and models, particularly in the field of ML, remains relatively understudied. As a prevailing practice of reusing datasets and models in the ML community could reinforce existing bias and stereotypes (Park & Li, 2025; Park & Cordell, 2023b) already reflected in training datasets, the study on data sharing characteristics calls attention from information scientists to develop a better version of the data lifecycle for ML beyond archival metadata (Dobreski et al., 2020). Further, current scholarship on data reuse explored the role of trust (Yoon, 2014; Yoon & Lee, 2019) in data reuse, especially in collaborative work (Al-Ani & Redmiles, 2009), perceived benefit (Joo et al., 2017) as well as community and career benefits (Yoon & Kim, 2020), but it is not clear how those findings are relevant to ML dataset reuse. To address this gap, further research is needed on how ML datasets are reused, in what contexts, with what perceived benefits, and their key characteristics. As a starting point, this study presents early work on descriptive characteristics of datasets and model sharing practices in Hugging Face. Hugging Face is particularly well-known among developer communities for its support of Transformer architectures, which are integrated into its pipelines (Wolf et al., 2020) with easy plug-and-play Python programming language. Gaining popularity in the developer community led to exponential growth of datasets (Yang et al., 2024) along with the tendency to reduce the cost of developing the models from scratch (Jiang, 2022). In parallel with the idea of Communities of Practice (CoP); how group of people share collective identity through consensual knowledge (Van House et al., 1998), Hugging Face as a developer community (a group of people publishing models and datasets through an open dataset sharing platform) provides an opportunity to empirically study data reuse practices in the computing field at scale. Specifically, we ask the following research questions: (RQ 1) What is the degree of diversity of datasets in training NLP models and CV models? (RQ 2) What is the degree of concentration of NLP datasets and CV datasets? (RQ 3) What is the degree of concentration of NLP models and CV models?

METHODOLOGY

Data collection

The datasets and models on NLP and CV tasks published on Hugging Face were collected on February 15th, 2025 through the Hugging Face API. A total of 28,434 NLP datasets and 397,595 NLP models from 12 NLP tasks (Question-answering, Text-generation, Feature-extraction, Table-question-answering, Sentence-similarity, Translation, Summarization, Zero-shot-classification, Text-classification, Token-classification, Text2text-generation, and Fill-mask) were collected, and 9,845 CV datasets and 79,596 CV models from 18 CV tasks (Mask-generation, Text-to-video, Image-to-video, Zero-shot-image-classification, Unconditional-image-generation, Image-to-text, Image-to-image, Image-classification, Text-to-image, Image-feature-extraction, Text-to-3D, Depth-

estimation, Zero-shot-object-detection, Keypoint-detection, Object-detection, Video-classification, Image-segmentation, and Image-to-3D) were collected.

Datasets and models reuse

Users of Hugging Face are required to provide self-reported metadata and a plain text description to support the management of models and datasets on the platform. Metadata records the base model if the model is a fine-tune, an adapter, or a quantized version (https://huggingface.co/docs/hub/model-cards), and this feature allows structuring hierarchy with parent (base) models and child (derivative) models. API provides access to information about the base model and can be further structured to examine which dataset was reused the most or which model was reused the most to develop derivative models in NLP and CV tasks. In answering RQ 1, we counted the number of datasets reused to train the models in NLP and CV tasks. For RQ 2 and RQ 3, we counted the number of times that the model and the dataset were reused as a base model and dataset, respectively. In our commitment toward supporting open, reproducible, and transparent data science, we share the code used in the study at https://github.com/park-jay/huggingface-models.

FINDINGS

In answering RQ 1, we focused on examining the diversity of datasets used to train NLP models and CV models. While NLP models for most tasks were trained on a little more than one dataset on average, the Zero-shot-classification task required the largest number of training datasets (μ=8.7574, σ=42.6210), followed by Sentence-similarity (μ=3.3647, σ=5.3457), Feature-extraction (μ=2.5782, σ=4.3964), and Text-generation (μ=2.3714, σ=4.27736). This finding indicates that the Zero-shot classification, Sentence similarity, Feature extraction, and Text generation synthesized diverse datasets to develop the model, compared to other tasks. Contrary to the observation of diverse dataset reuse in the Zero-shot-classification, Sentence-similarity, Feature-extraction, and Text-generation in NLP tasks, all CV models were trained on fewer than two datasets on average. The largest number of training datasets used for the model is Image-feature-extraction (μ=1.7600, σ=1.9579), followed by Depth-estimation (μ=1.7500, σ=2.3011), Video-classification (μ=1.6666, σ=0.9759), and Image-to-text (μ=1.6422, σ=2.9112).

To address RQ 2, the number of reuses of NLP and CV datasets was counted. Overall, many NLP task datasets remained less than one time, indicating that even though the dataset is available through the Hugging Face platform, it has not been widely used by the ML community. Table-question-answering (μ=0.0274, σ=0.2516), Translation (μ=0.1171, σ=0.6341), and Zero-shot-classification (μ=0.1244, σ=0.7998) remained the least three tasks reusing NLP datasets. The NLP datasets that had more than one reuse on average are found in Text-generation (μ=1.8877, σ=18.1683) and Sentence-similarity (μ=1.8331, σ=11.0991) tasks. Text-generation and Sentence-similarity datasets had high standard deviation as well, which may indicate that the dataset reuse may reflect a wide gap between the popularity of datasets. The popular datasets get an opportunity to be reused, while unpopular datasets do not have an opportunity to be reused (Zipf, 1949). Contrary to NLP, we observed more frequent dataset reuse in CV tasks. The highest CV task that had the dataset reuse is text-to-image (μ=0.6354, σ=0.5373), followed by Object-detection (μ=0.2351, σ=0.2522) and Image-classification (μ=0.2084, σ=0.2974).

For RQ 3, we examined how frequently existing models are reused in both the NLP and CV domains. While the majority of NLP models were reused between one time or two times on average, similar to the NLP task that reused datasets the most, the Text-generation models reused other models the most (μ=1.3347, σ=1.5012) on average. The Summarization model (μ=1.1769, σ=1.9735) and Zero-shot-classification model (μ=1.1417, σ=0.7373) followed the Text-generation models. High model reuse in the Text-generation task may indicate that the ML community is concentrating on creating GPT-like language models and shows dependency on existing models. In the CV domain, model reuse appears more frequent. Among CV tasks, Unconditional-image-generation models (μ=3.2500, σ=4.5000) re(used) other models on average, followed by Image-to-3d (μ=2.0000, σ=3.0000), and Image-to-video models (μ=1.3636, σ=0.9021). Overall, the differences in datasets reuse (RQ 2) and model reuse (RQ 3) patterns may suggest varied ML community norms and practices.

CONCLUSION

In conclusion, this study explores distinct patterns in the reuse of datasets and models across NLP and CV tasks in Hugging Face. NLP tasks, such as Zero-shot classification, Sentence-similarity, Feature-extraction, and Text-generation, used diverse datasets to train models, while CV models relied on fewer (and less diverse) datasets to train models (RQ 1). NLP Datasets in Text-generation and Sentence-similarity were reused the most, indicating high reliance on existing datasets, while CV datasets were comparably less reused (RQ2). For RQ 3, we found CV models reuse more compared to NLP models. While we yet do not have any explanation for the findings at this point, building on factors identified from previous studies, such as the concept of trust, documentation, and community validation (Yoon & Lee, 2019), our next goal is to explore the developers' motivation behind the differences between NLP and CV data reuse.

GENERATIVE AI USE

We used Copilot, available through Visual Studio Code, to write code for data collection and analysis.

AUTHOR ATTRIBUTION

Jaihyun Park: Conceptualization (lead), Formal Analysis (lead), Methodology, Writing (Original Draft Preparation).
Ayoung Yoon: Conceptualization (supporting), Investigation, Validation, Writing (Original Draft Preparation).

ACKNOWLEDGMENTS

Jaihyun Park gratefully acknowledges the support from the Eugene Garfield Doctoral Dissertation Fellowship 2024.

REFERENCES

Al-Ani, B., & Redmiles, D. (2009, July). In strangers we trust? Findings of an empirical study of distributed teams. In *2009 Fourth IEEE International Conference on Global Software Engineering* (pp. 121-130). IEEE.

Dobreski, B., Park, J., Leathers, A., & Qin, J. (2020, March). Remodeling archival metadata descriptions for linked archives. In *Proceedings of the international conference on dublin core and metadata applications* (pp. 1-11).

Huang, S., & Goh, D. H. (2024). Public Perceptions of GAI: A YouTube Analysis of Sora. *Proceedings of the Association for Information Science and Technology*, *61*(1), 940–942. https://doi.org/10.1002/pra2.1147.

Jiang, W., Synovic, N., Sethi, R., Indarapu, A., Hyatt, M., Schorlemmer, T. R., ... & Davis, J. C. (2022, November). An empirical study of artifacts and security risks in the pre-trained model supply chain. In *Proceedings of the 2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses* (pp. 105-114).

Joo, S., Kim, S., & Kim, Y. (2017). An exploratory study of health scientists' data reuse behaviors: Examining attitudinal, social, and resource factors. *Aslib Journal of Information Management*, *69*(4), 389-407.

Park, J., & Cordell, R. (2023a, December). A Quantitative Discourse Analysis of Asian Workers in the US Historical Newspapers. In *The Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages* (p. 7).

Park, J., & Cordell, R. (2023b). The ripple effect of dataset reuse: Contextualising the data lifecycle for machine learning data sets and social impact. *Journal of Information Science*, 01655515231212977.

Park, J., & Cordell, R. (2025, May). A Data-driven Investigation of Euphemistic Language: Comparing the usage of "slave" and "servant" in 19th century US newspapers. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities* (pp. 350-364).

Park, J., & Jeoung, S. (2022, May). Raison d'être of the benchmark dataset: A survey of current practices of benchmark dataset sharing platforms. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP* (pp. 1-10).

Park, J., & Li, K. (2025). Rethinking Reuse in Data Lifecycle in the Age of Large Language Models. *Information Matters*, *5*(4).

Park, J., Yang, J., Tolbert, A., & Bunsold, K. (2024). You change the way you talk: Examining the network, toxicity and discourse of cross-platform users on Twitter and Parler during the 2020 US Presidential Election. *Journal of Information Science*, 01655515241238405.

Griebel, S., Cohen, B., Li, L., Park, J., Liu, J., Perkins, J., & Underwood, T. (2024, December). Locating the Leading Edge of Cultural Change. In *CHR 2024: Computational Humanities Research Conference* (pp. 232-245).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Yang, X., Liang, W., & Zou, J (2024). Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on HuggingFace. In *The Twelfth International Conference on Learning Representations*.

Yoon, A. (2014). End users' trust in data repositories: definition and influences on trust development. *Archival Science*, *14*(1), 17-34.

Yoon, A., & Kim, Y. (2020). The role of data-reuse experience in biological scientists' data sharing: an empirical analysis. *The Electronic Library*, *38*(1), 186-208.

Yoon, A., & Lee, Y. Y. (2019). Factors of trust in data reuse. *Online Information Review*, *43*(7), 1245-1262.

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998, November). Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work* (pp. 335-343).

Zhou, C., Kuttal, S. K., & Ahmed, I. (2018, October). What makes a good developer? an empirical study of developers' technical and social competencies. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 319-321). IEEE.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.