Characteristics of Dataset and Model Reuse Patterns on the Hugging Face Platform

Jaihyun Park, Ph.D.
Ayoung Yoon, Ph.D.

Department of Library and Information Science
Luddy School of Informatics, Computing, and Engineering
Indiana University Indianapolis

INTRODUCTION

Hugging Face

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-E3). 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.

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 biases and stereotypes (Park & Cordell, 2023) 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.

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 the exponential growth of datasets (Yang et al., 2024) along with the tendency to reduce the cost of developing the models from scratch (Jiang, 2022).

As a starting point, this study presents early work on descriptive characteristics of datasets and model sharing practice in Hugging Face. In parallel with the idea of Communities of Practice (CoP); how a group of people share collective identity through consensual knowledge (Van House et al., 1998), Hugging Face as a developer community 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

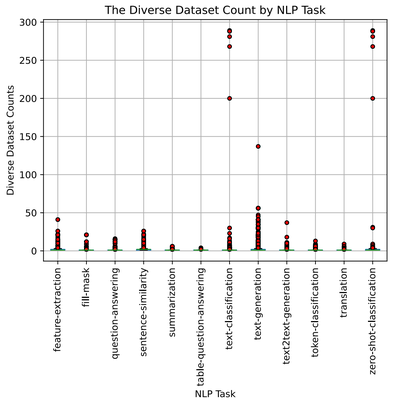
Date: February 15th, 2025
NLP datasets: n=28,434
NLP models: n=397,595
CV datasets: n=9,845
CV models: n=79,596

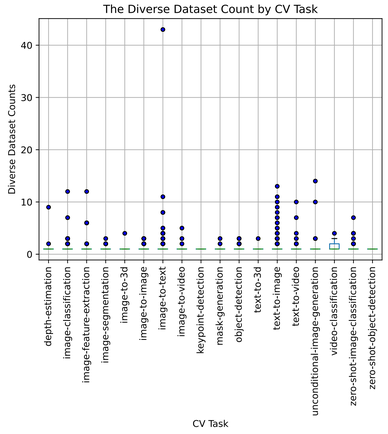
Datasets and model reuse

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, 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 dataset and model, respectively.

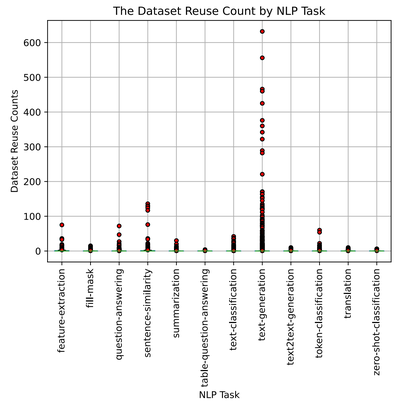
FINDINGS

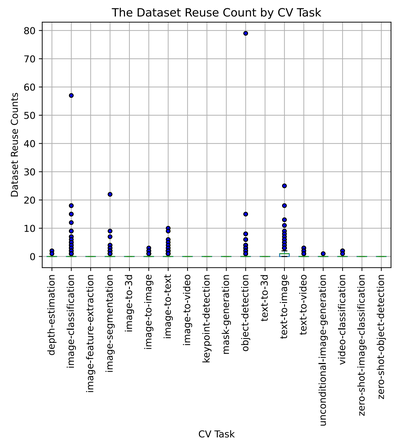




In answering RQ 1, we found that NLP models for most tasks were trained on a little more than one dataset on average, **Zero-shot classification** (μ=8.7574, σ=42.6210), followed by **Sentence-similarity** (μ=3.3647, σ=5.3457), **Feature-extraction** (μ=2.5782, σ=4.3964). All CV models were trained on 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), and **Video-classification** (μ=1.6666, σ=0.9759).

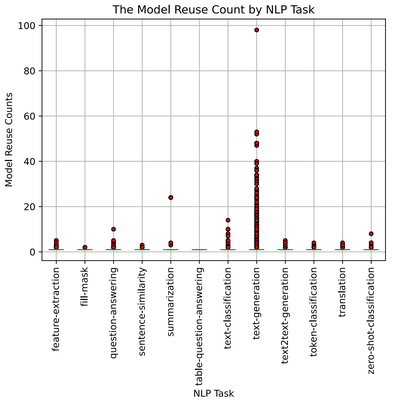


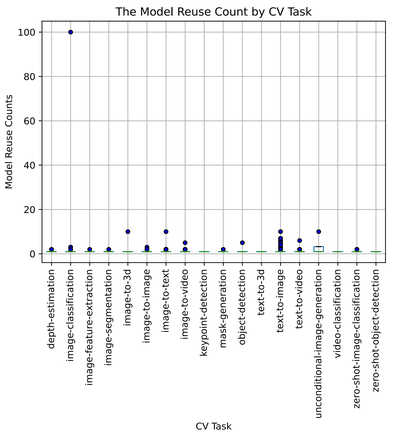


For RQ 2, we found that many NLP task datasets remained less than one time, indicating that even though the dataset is available through the Hugging Face platform, it has not been widely used in the ML community. **Text-generation** (μ=1.8877, σ=18.1683) and **Sentence-similarity** (μ=1.8331, σ=11.0991) had more than one reuse on average. The highest CV task that had the dataset reuse is **Text-to-image** (μ=0.6354, σ=0.5373), **Object-detection** (μ=0.2351, σ=0.2522), and **Image-classification** (μ=0.2084, σ=0.2974).

For RQ 3, we found Text-generation reused other NLP models the most (μ=1.3347, σ=1.5012), followed by **Summarization** (μ=1.1769, σ=1.9735), and **Zero-shot-classification** (μ1.1417,= σ=1.5012). For CV tasks, **Unconditional-image-generation** (μ=3.2500, σ=4.5000) followed by **Image-to-3d** (μ=2.0000, σ=3.0000), and **Image-to-video** (μ=1.3636, σ=0.9021).

Through answering RQs, we found the general trend that the popular datasets and models get an opportunity to be reused, while unpopular datasets do not have an opportunity to be reused (Zipf, 1949).





ACKNOWLEDGEMENTS

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

REFERENCES

Park J., & Cordell, R. (2023). 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., & 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).

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*.

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).

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).

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

LUDDY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING
Indianapolis