COMP 8977 Internship Project II

University of Windsor, School of Computer Science

Project Proposal

Project Title: On the Popularity of Hugging Face Dataset

Term: Fall 2025

Student Names & IDs:

Team 4:

Dhruvil Chauhan – 110183580

Vivek Savalia – 110159925

Supervisor Name: Dr. Muhammad Asaduzzaman

Introduction

Hugging Face has quickly grown into one of the biggest and most widely used open-source platforms for AI. It's more than just a tool it's become a community space where people can find and share datasets and pre-trained models. This openness has made it much easier for researchers, developers, and even beginners to access and build on cutting-edge AI without starting from scratch.

On the dataset side, Hugging Face isn't limited to text anymore it now supports computer vision, audio, and even multimodal datasets. Many of these have gone on to influence not just academic research, but also real-world applications. On the model side, it gives people access to powerful pre-trained models like BERT, GPT, and LLaMA, which continue to push the boundaries of what's possible with AI.

But there's a catch. Not all contributions get equal attention. While some models and datasets attract millions of downloads and huge community engagement, others despite being high quality end up overlooked. This uneven popularity raises important questions: What makes a dataset or model gain traction? Why do some resources rise to the top while others fade into the background?

This project aims to dig into those questions. We want to study Hugging Face datasets and models systematically to understand the key factors that drive popularity. By looking at things like metadata, documentation, technical features, and community interaction, we will try to identify patterns and signals that matter most.

The goal is to provide practical insights for creators who want their work to be seen, for developers and researchers looking for the right tools, and even for the Hugging Face platform itself to improve discoverability. In the end, this study could help ensure that great contributions don't go unnoticed and that the community benefits more evenly.

Motivation

The motivation behind this project comes from a mix of academic, practical, and community-driven needs.

First, from a research point of view, the datasets and models that become popular often end up shaping the direction of entire fields. Researchers rely on these as benchmarks, so it's important to understand what makes them widely adopted. If we can identify the factors that drive popularity, we can make sure that strong, credible resources set the standard for future work.

For developers and contributors, visibility is everything. A high-quality dataset or model isn't useful if nobody can find or use it. Right now, the guidelines for making contributions stand out aren't very clear. By analyzing things like metadata, documentation, and technical details, we hope to create data-backed recommendations that will help authors improve discoverability and reach more users.

On a community level, Hugging Face is powered by collaboration. The more accessible and reusable models and datasets are, the faster the whole AI ecosystem grows. Well-documented and easy-to-use resources encourage others to build, share, and push the boundaries of what's possible.

Finally, there's an educational angle. Many students and newcomers take their first steps in AI through Hugging Face. By highlighting what makes certain resources more usable and popular, we can give them a clearer roadmap to follow, support reproducibility, and reinforce open science practices.

Background Study

The idea of studying popularity in open-source platforms isn't entirely new. For example, Borges et al. (2016) found that factors like good documentation and frequent updates were strongly linked to how popular GitHub repositories became. Similarly, Zhu et al. (2014) showed that something as basic as how files are organized can affect whether a project gets reused.

More recently, researchers have started to turn their attention to Hugging Face itself. Jiang et al. (2023) explored how models are reused and adapted. Liang et al. (2024) looked at more than 32,000 model cards to see how documentation quality varies. Pepe et al. (2024) studied datasets, focusing on bias, licensing, and how complete the metadata is.

But here's the gap: no study so far has looked at both datasets and models together to understand what really drives popularity on Hugging Face. That's where our project comes in. Hugging Face offers a wealth of metadata model cards, licenses, tags, downloads, likes, and more that make large-scale analysis possible. By examining these features across both datasets and models, we aim to spot meaningful patterns, uncover correlations, and even build predictive models that can forecast popularity.

Project Goals

The main goal of this project is to understand what makes certain Hugging Face models and datasets more popular than others, and to translate those insights into practical guidance for the community.

More specifically, the project aims to:

- Identify key features – Analyze repository-level details and metadata characteristics that are linked to popularity.
- Build predictive models – Develop classifiers that can distinguish popular from less popular resources based on documentation quality, structure, platform integration, and technical attributes.
- Rank the most important factors – Use interpretable methods such as permutation importance and SHAP to highlight which features have the strongest influence on popularity.
- Test generalizability – Check whether the findings hold true across domains (NLP, Computer Vision, Audio, Multimodal) and across different affiliations (companies, universities, individuals, and community contributors).
- Provide actionable guidelines – Develop best practices for model and dataset creators to improve visibility, adoption, and reproducibility on Hugging Face.

Beyond models, this project will also extend to datasets, which face similar challenges in visibility and adoption. By collecting and analyzing all available datasets on the Hugging Face Hub using the same methodology, we will be able to compare popularity drivers across models and datasets. This broader scope will lead to a unified understanding of how resources gain traction on the platform.

Methodology

Non-Technical Overview

At a high level, our approach is designed to stay lightweight we'll only work with metadata from Hugging Face, not the heavy model weights or dataset files. Here's the plan:

- Collect metadata using the Hugging Face API.
- Clean and standardize the fields so thousands of entries can be compared consistently.
- Look for patterns by exploring correlations between different metadata features and popularity signals (downloads, likes, etc.).
- Build simple, interpretable models (like linear models, decision trees, and SVMs) to figure out which factors matter most.
- Communicate results clearly through easy-to-read visuals and plain-English recommendations for dataset/model creators.

Technical Steps

On the technical side, the process will involve several key stages:

- Data Acquisition – Collect a snapshot of Hugging Face metadata through the API.
- Feature Engineering – Extract and structure useful features such as README/documentation length, presence of BibTeX, tags, and technical scope.
- Statistical Analysis – Use tests like Mann-Whitney U and Chi-square to check for significant differences between popular vs. less popular entries.
- Predictive Modeling – Train models including Random Forests, Decision Trees, and SVMs to predict popularity and estimate the relative importance of features.
- Interpretability – Apply SHAP values and permutation importance to make the results transparent and explain why certain features influence popularity

Roadmap of the Project

The project spans nine weeks, with each week dedicated to a specific phase of work.

Week 1 (Sep 28 – Oct 4):

- Finalize research questions, define popularity metrics, and feature taxonomy.
- Set up repository, project board, and coding standards.

Week 2 (Oct 5 – Oct 11):

- Implement metadata collector using the Hugging Face API with rate limiting and retries.
- Prototype unified schema and sample 1–2k datasets for feasibility testing.

Week 3 (Oct 12 – Oct 18):

- Perform data cleaning and normalization, addressing missing and heterogeneous fields.
- Conduct exploratory data analysis (EDA): distributions, correlations, and modality stratification.

Week 4 (Oct 19 – Oct 25):

- Carry out feature engineering: documentation richness, update recency, affiliation, modality and language flags, and size categories.
- Define training and evaluation splits and run baseline models for feature importance (linear and tree-based).

Week 5 (Oct 26 – Nov 1):

- Refine models with SHAP and permutation importance.
- Apply cross-validation and robustness checks.
- Conduct error analysis and guard against confounders (e.g., dataset age).

Week 6 (Nov 2 – Nov 8):

- Perform ablation studies by feature family and modality, plus sensitivity analyses.
- Draft preliminary findings and recommendations.

Week 7 (Nov 9 – Nov 15):

- Build a lightweight dashboard (Streamlit/Plotly) to explore key drivers of popularity.
- Conduct peer review of methodology and results and integrate feedback.

Week 8 (Nov 16 – Nov 22):

- Finalize the report narrative, polish visuals, and compile references.
- Prepare a creator checklist and best-practice guidance.

Week 9 (Nov 23 – Nov 29):

- Complete final QA and reproducibility checks and archive artifacts.
- Prepare slides and demo, then submit all final materials.

Resources

- Skills: Python, API usage, data wrangling, ML fundamentals, visualization.
- Libraries: Pandas, NumPy, Scikit-learn, SHAP, Matplotlib, Seaborn, Plotly.
- Tools: Hugging Face Hub API, GitHub Projects, Streamlit/Dash, Trello.

References

- Wolf et al. (2020). Transformers: State-of-the-Art NLP.
- Gebru et al. (2021). Datasheets for Datasets. CACM.
- Mitchell et al. (2019). Model Cards for Model Reporting. FAT*.
- Borges et al. (2016). Understanding GitHub Popularity. IEEE ICSME.
- Jiang et al. (2023). Reuse in Hugging Face. IEEE ICSE.
- Liang et al. (2024). Systematic Analysis of Model Cards. Nat. Mach. Intell.
- Pepe et al. (2024). Bias & Licenses in Hugging Face Repositories. ICPC.