

Towards a Classification of Open-Source ML Models and Datasets for Software Engineering

Alexandra González
Universitat Politècnica de Catalunya
Barcelona, Spain
alexandra.gonzalez.alvarez@upc.edu

Xavier Franch
Universitat Politècnica de Catalunya
Barcelona, Spain
xavier.franch@upc.edu

David Lo
Singapore Management University
Singapore
silverio.martinez@upc.edu

Silverio Martínez-Fernández
Universitat Politècnica de Catalunya
Barcelona, Spain
silverio.martinez@upc.edu

Abstract—Background: Open-Source Pre-Trained Models (PTMs) and datasets provide extensive resources for various Machine Learning (ML) tasks, yet these resources lack a classification tailored to Software Engineering (SE) needs. **Aims:** We apply an SE-oriented classification to PTMs and datasets on a popular open-source ML repository, Hugging Face (HF), and analyze the evolution of PTMs over time. **Method:** We conducted a repository mining study. We started with a systematically gathered database of PTMs and datasets from the HF API. Our selection was refined by analyzing model and dataset cards and metadata, such as tags, and confirming SE relevance using Gemini 1.5 Pro. All analyses are replicable, with a publicly accessible replication package. **Results:** The most common SE task among PTMs and datasets is *code generation*, with a primary focus on *software development* and limited attention to *software management*. Popular PTMs and datasets mainly target *software development*. Among ML tasks, *text generation* is the most common in SE PTMs and datasets. There has been a marked increase in PTMs for SE since 2023 Q2. **Conclusions:** This study underscores the need for broader task coverage to enhance the integration of ML within SE practices.

Index Terms—Pre-trained models for Software engineering, Software engineering datasets, Hugging Face

I. INTRODUCTION

The fast expansion of open-source platforms like Hugging Face (HF) [1] has enhanced access to Machine Learning (ML) models and datasets, driving advancements across various domains. With a consistent and significant uptrend in development activities on HF [2], it is distinguished by its vast collection of Pre-Trained Models (PTMs), compared to other platforms [3] [4]. However, the categorization of these resources overlooks the specific needs of Software Engineering (SE). SE tasks frequently involve *code generation*, *code analysis*, and *bug detection*, which differ significantly from the tasks commonly addressed by general-purpose ML models such as *object detection* or *image segmentation*. Therefore, the motivation for this work is to address this gap, as the absence of SE-specific categorization limits the efficient application of ML in SE tasks, potentially slowing down SE innovation. By providing a framework that aligns ML tasks with SE

needs, this research aims to make the selection of PTMs and datasets more relevant and effective for SE practitioners and researchers, thus addressing a critical need within the field [5].

The main contributions of this work are: (a) proposing and proving the feasibility of a preliminary classification framework for PTMs and datasets hosted on HF, tailored to SE needs; (b) providing advanced analysis, including the exploration of the relationship between SE activities and ML tasks, as well as the evolution of SE PTMs over time; (c) presenting a reproducible pipeline that accesses the HF API, filters, refines, and classifies resources on specific SE tasks.

Data availability statement: All research components, including the original and preprocessed data, along with all scripts for data collection, preparation, and analysis, are publicly available on Zenodo [6]. This ensures transparency and enables independent replication of the study, which is essential for updating the classification as new open-source PTMs and datasets are constantly being released.

II. RELATED WORK

A systematic literature review conducted by Hou et al. [7] analyzed 395 research papers from January 2017 to January 2024 and categorized Large Language Models (LLMs) into SE tasks. These tasks were grouped into SE activities according to the six phases of the **Software Development Life Cycle**: *requirements engineering*, *software design*, *software development*, *software quality assurance*, *software maintenance*, and *software management*. Di Sipio et al. [5] highlighted the lack of a SE classification of PTMs on HF, as the existing one is specific for ML. To address this gap, they proposed extracting information from the model cards [8] and using a semi-automated method to identify SE tasks and their corresponding PTMs from the literature. However, they only tested the mapping on three PTMs: *BERT*, *RoBERTa*, and *T5*. Yang et al. [9] analyzed the ecosystem of LLMs as of August 2023, curating 366 models and 73 datasets from HF for SE. The study also explores the use of LLMs to assist in constructing and analyzing the ecosystem, which increased the model size

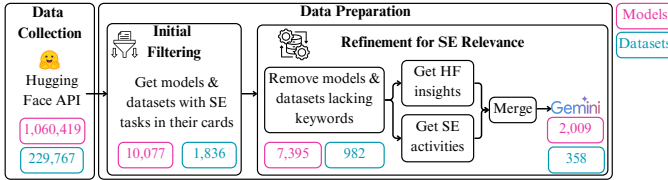


Fig. 1: Data collection and preparation pipeline.

by 16.5. They focus on code-based LLMs, which represent a more specific scope compared to LLMs for broader SE tasks.

Despite the above notable advances, a comprehensive framework that organizes PTMs and datasets on HF with a strong SE orientation remains absent. In contrast to prior works, this paper adopts a SE perspective to extend the existing taxonomy in Hou et al. [7], to encompass a broader and more diverse set of PTMs, as well as datasets, categorizing them according to specific SE tasks and activities. By analyzing a substantially larger set of PTMs and SE-relevant datasets on HF, we address the unique SE-specific requirements unmet by previous studies, offering a novel, exhaustive preliminary classification framework for the community.

III. METHODOLOGY

A. Research Objectives

Following the Goal Question Metric (GQM) template [10], our goal is to **analyze PTMs and datasets for the purpose of their classification with respect to their application to SE tasks and activities from the point of view of software engineers in the context of the HF Hub**.

This goal is structured around four RQs. First, we need to assess the quality of model and dataset cards, as this information is essential for the subsequent RQs:

- **RQ1:** What is the status of the model and dataset cards?

Next, we explore how the selected resources tackle SE:

- **RQ2:** How do PTMs and datasets address SE?
 - **RQ2.1:** What SE tasks and activities are covered by PTMs and datasets?
 - **RQ2.2:** What are the most popular PTMs and datasets that address SE activities?

Following this, we explore the connection between ML tasks in HF’s classification and SE activities:

- **RQ3:** How are ML tasks related to SE activities?

Lastly, we examine the long-term relevance of our findings:

- **RQ4:** How stable is this information over time?

B. Data collection and Preparation

Figure 1 illustrates the pipeline we followed to collect and prepare data for analysis. This process is designed for reproducibility, allowing anyone with access to the replication package [6] to validate and update the results. The inclusion criteria are summarized in Table I.

TABLE I: Inclusion criteria.

- 1) PTMs and datasets must be available on HF.
- 2) PTMs and datasets must contain valid model and dataset cards.
- 3) The cards must specify at least one SE task [7].
- 4) The cards must include “code” or “software”.
- 5) An LLM (Gemini 1.5 Pro) must confirm relevance to SE.

1) **Data Collection:** We used the HF API [11] to gather all available resources as of October 19, 2024: 1,060,419 PTMs and 229,767 datasets. During this process, we collected the unique identifiers for each model and dataset, and assessed the availability of their cards.

2) Data Preparation:

a) **Initial Filtering:** We automatically filtered PTMs and datasets mentioning SE tasks proposed by Hou et al. [7] in their cards. This step enabled us to focus on SE-relevant resources, resulting in 10,077 PTMs and 1,836 datasets.

b) **Refinement for SE Relevance:** To ensure our focus on SE, we removed entries that did not contain “code” or “software” in the model and dataset cards, resulting in 7,395 PTMs and 982 datasets. For this subset, we retrieved all available information from HF and mapped the SE task to the corresponding SE activity. Furthermore, we used an LLM, Gemini 1.5 Pro [12], to identify whether each resource was intended for SE based on an analysis of their cards and metadata. As a validation step, we manually classified a balanced sample of 30 PTMs and 30 datasets (between SE-relevant and non-SE), and compared our decisions with those generated by the LLM using Cohen’s Kappa [13]. This sample size was chosen to ensure reliable initial validation, as higher levels of agreement are anticipated [14]. For PTMs, the Kappa was 0.80, which falls within the 0.61 to 0.80 range indicating substantial agreement [15]; for datasets, it was 0.70, also reflecting this level of agreement. We noticed that some SE tasks, such as *logging* and *verification*, could be ambiguous. For instance, a model card containing a code snippet like `from transformers import logging` might be misclassified as a *logging* task when it may not be. To ensure rigor, we asked the LLM to classify them by providing explicit definitions of such tasks. Lastly, we applied the prompts to all resources, resulting in 2,009 PTMs and 358 datasets, representing 0.19% and 0.16% of the original sets.

3) **Data Analysis:** In exploring the landscape of PTMs and datasets within HF, we centered our analysis on their cards, as well as on metadata such as tags, creation dates, and other relevant attributes. Additionally, we define popularity as the sum of the normalized number of likes and downloads.

IV. RESULTS

A. What is the status of the model and dataset cards? (RQ1)

As summarized in Table II, over 33% of the PTMs lack a model card, highlighting a gap in the documentation. Furthermore, more than 65% do not mention any SE tasks, and only 0.95% mention SE tasks. The analysis for datasets reveals a

TABLE II: Summary of model cards.

Category	Number of PTMs	Proportion
Not available	350,524	33.05%
Available but empty	3,686	0.35%
No SE tasks	696,132	65.65%
With SE tasks	10,077	0.95%

TABLE III: Summary of dataset cards.

Category	Number of datasets	Proportion
Not available	64,210	27.95%
Available but empty	1,294	0.56%
No SE tasks	162,427	70.69%
With SE tasks	1,836	0.80%

lack of documentation, as detailed in Table III. Over 27% of datasets do not have a dataset card, indicating a significant gap in available information. Additionally, 0.56% are empty, and a staggering 70.69% do not reference any SE task. Only 0.80% of the datasets allude to SE tasks.

Finding 1: The current state of HF shows that only 10,077 PTMs and 1,836 datasets mention SE tasks in their cards.

B. How do PTMs and datasets address SE? (RQ2)

1) **What SE tasks and activities are covered by PTMs and datasets?:** Figure 2 shows the distribution of PTMs across SE tasks, color-coded by SE activity. The top tasks are *code generation* and *code completion*, with the first having nearly ten times as many PTMs as the third most common. The most represented activity is *software development*, while *software design* and *requirements engineering* have limited PTMs. Additionally, the absence of PTMs for *software management* highlights a critical gap in the current landscape.

Regarding datasets, Figure 3 shows that *software development* dominates, while *software design* and *requirements engineering* are underrepresented, and *software management* is entirely absent, mirroring the patterns observed with PTMs.

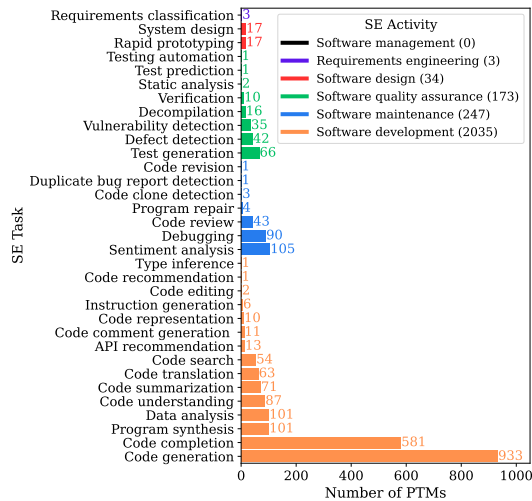


Fig. 2: PTMs associated with each SE task and SE activity.

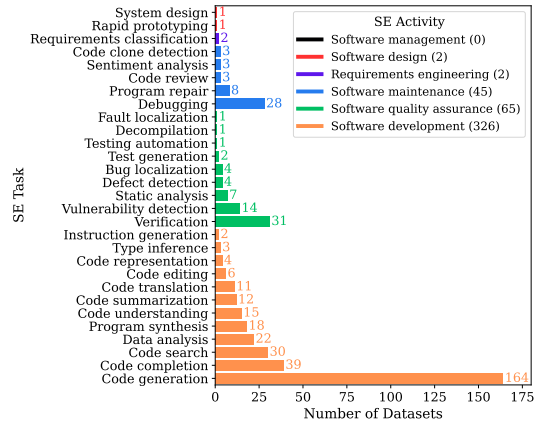


Fig. 3: Datasets associated with each SE task and SE activity.

The leading SE task is *code generation*, which has over three times the number of datasets compared to the next task.

Interestingly, the focus of secondary SE activity diverges between PTMs and datasets: *software maintenance* is the second SE activity most addressed for PTMs, while for datasets, *software quality assurance* holds this position.

Finding 2.1: *Code generation* is the most covered SE task among PTMs and datasets.

Finding 2.2: *Software development* dominates, while *software management* is absent in PTMs and datasets.

2) **What are the most popular PTMs and datasets that address SE activities?:** As shown in Figure 4, the three most popular PTMs across SE activities reveal that *software development* has a higher popularity due to its broad representation. Figure 5 presents the three most popular datasets for *software development*, *software quality assurance*, and *software maintenance*. Other SE activities are omitted, as their dataset popularity is exactly 0. Popularity values for datasets tend to be higher, which may suggest that the community places more value on datasets than on PTMs, potentially indicating a higher demand for quality datasets in comparison to PTMs. In other words, researchers might perceive a good dataset as more valuable or essential for advancing SE activities than a well-performing model.

Finding 2.3: The most popular PTMs and datasets predominantly address *software development*.

Finding 2.4: Datasets tend to be more popular than PTMs, suggesting a higher demand for quality datasets in SE.

C. How are ML tasks related to SE activities? (RQ3)

The relationship between SE activities and ML tasks for PTMs and datasets is illustrated in Figures 6 and 7, respectively, with the flow indicating the number of resources. Both figures highlight the prominence of the ML task *text generation*, which is associated with the largest number of PTMs and datasets. This task is most commonly linked to *software development* from the SE perspective, though it also

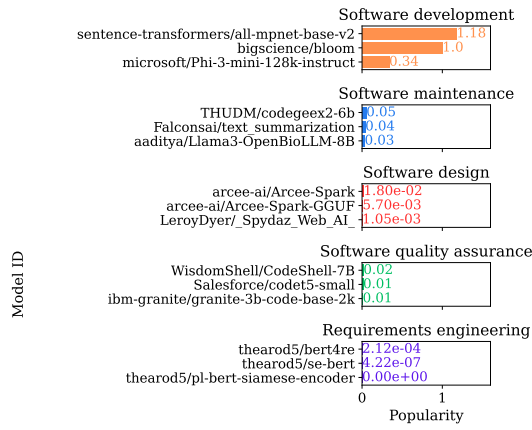


Fig. 4: Top 3 most popular PTMs per SE activity.

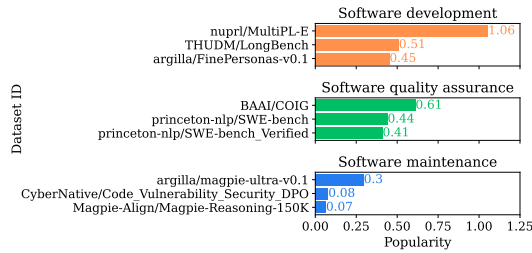


Fig. 5: Top 3 most popular datasets per SE activity.

supports other SE activities. In contrast, other ML tasks exhibit considerably smaller flows.

Finding 3: The ML task *text generation* is the most common among SE-related PTMs and datasets.

D. How stable is this information over time? (RQ4)

Figure 8 shows a growth in the creation of PTMs for SE tasks since 2020, with quarterly data revealing periods of accelerated development. Notably, the percentage of PTMs for *software development* increased from 6.78% in 2023 Q2 to 20.55% currently. Similarly, *software maintenance* and *software quality assurance* also showed significant growth, with the former rising from 1.70% to 32.34%, and the latter increasing from 1.19% to 28.57% over the same period. A zoom-in view of this period, from 2023 Q2 to 2024 Q3, is

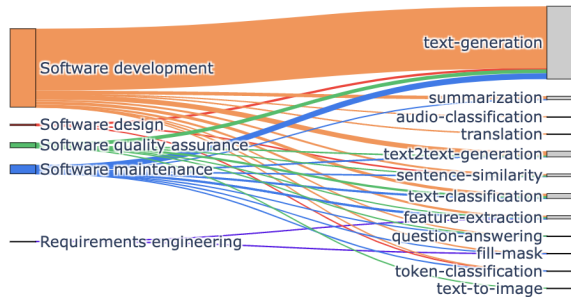


Fig. 6: Association of SE activities with ML tasks for PTMs.

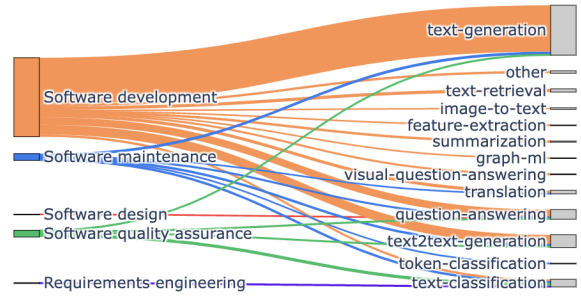


Fig. 7: Association of SE activities with ML tasks for datasets.

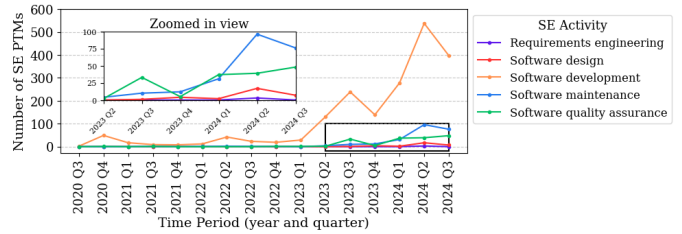


Fig. 8: Number of SE PTMs created since 2020.

included in the figure to provide a closer examination. While the ranking of activities remains relatively stable over time, there are some fluctuations, particularly in 2023 Q3 and 2024 Q1, where *software quality assurance* outperformed *software maintenance*, albeit with a less pronounced change in 2024.

Finding 4.1: PTMs for SE have grown consistently since 2020, with notable acceleration from 2023 Q2.

Finding 4.2: The ranking of SE tasks remains stable over time, with minor fluctuations.

V. DISCUSSION AND LIMITATIONS

Our results align with Hou et al. [7], which shows a similar distribution of LLMs across SE activities, with both studies identifying an emphasis on *code generation* tasks. However, we have found a gap in *software maintenance* within the current state of HF. Notably, our findings complement Yang et al.’s analysis [9], highlighting a rapid growth in this area.

Potential threats impacting our study’s validity are outlined. First, the quality of the model and dataset cards may compromise internal validity, as incomplete documentation could lead to misclassification. Our conclusions rely on the assumption that each PTM and dataset’s card accurately reflects the SE tasks it addresses. Second, other platforms/repositories (e.g., PyTorch Hub [16]) may host additional relevant resources, and changes on HF could affect the broader applicability of our conclusions. Third, our classification of SE tasks and activities is based on a taxonomy from the existing literature [7]. Although this taxonomy informs our analysis, emerging standards may further enrich our understanding of SE. To mitigate these threats, our study is fully replicable, enabling future researchers to validate and extend our findings.

VI. CONCLUSION AND FUTURE WORK

This study examines the availability of PTMs and datasets for SE hosted on HF, revealing that almost 1% reference SE tasks. PTMs predominantly focus on *code generation* and *code completion*, while *code generation* is the most represented dataset task. SE activities like *software design* and *requirements engineering* are under-represented, with a gap in *software management* resources. SE datasets tend to be more popular than PTMs. The most prevalent ML task across both PTMs and datasets is *text generation*. Additionally, there has been a significant surge in PTMs for SE since 2023 Q2. This snapshot of the current landscape highlights the existing gaps in SE resources and provides valuable insight into the field's evolving needs, helping to motivate future research efforts aimed at addressing these underrepresented areas.

For future work, we plan to extend this classification to other repositories, such as Papers with Code [17], PyTorch Hub [16] and TensorFlow Hub [18], allowing for a broader analysis of PTMs and datasets in SE. In addition, we aim to refine the classification process by addressing current limitations, such as handling synonymous in SE tasks, to improve accuracy. To make our classification more actionable, we propose developing a dashboard to assist researchers with proper sampling and support practitioners in the SE community when selecting PTMs for real-world applications, thereby enhancing the relevance and impact of our work.

ACKNOWLEDGMENT

This work has been funded by the Spanish research project DOGO4ML (ref. PID202 0-117191RB-I00).

REFERENCES

- [1] Hugging Face Inc., “Hugging Face – The AI community building the future. — huggingface.co,” <https://huggingface.co>, [Accessed: 28-10-2024].
- [2] J. Castaño, S. Martínez-Fernández, X. Franch, and J. Bogner, “Analyzing the evolution and maintenance of ml models on hugging face,” in *Proceedings of the 21st International Conference on Mining Software Repositories*, ser. MSR '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 607–618. [Online]. Available: <https://doi.org/10.1145/3643991.3644898>
- [3] L. Gong, J. Zhang, M. Wei, H. Zhang, and Z. Huang, “What is the intended usage context of this model? an exploratory study of pre-trained models on various model repositories,” *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 3, pp. 1–57, 2023.
- [4] W. Jiang, Y. Yasmin, J. Jones, N. Synovic, J. Kuo, N. Bielanski, Y. Tian, G. K. Thiruvathukal, and J. C. Davis, “Peatmoss: A dataset and initial analysis of pre-trained models in open-source software,” in *Proceedings of the 21st International Conference on Mining Software Repositories*, ser. MSR '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 431–443. [Online]. Available: <https://doi.org/10.1145/3643991.3644907>
- [5] C. Di Sipio, R. Rubei, J. Di Rocco, D. Di Ruscio, and P. T. Nguyen, “Automated categorization of pre-trained models in software engineering: A case study with a Hugging Face dataset,” in *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 351–356. [Online]. Available: <https://doi.org/10.1145/3661167.3661215>
- [6] Anonymous, “Replication package for “classifying open-source pre-trained models and datasets for software engineering,”” Nov. 2024. [Online]. Available: <https://zenodo.org/records/14057757>
- [7] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, and H. Wang, “Large Language Models for Software Engineering: A Systematic Literature Review,” *ACM Trans. Softw. Eng. Methodol.*, Sep. 2024, just Accepted. [Online]. Available: <https://doi.org/10.1145/3695988>
- [8] 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*, 2019, pp. 220–229.
- [9] Z. Yang, J. Shi, P. Devanbu, and D. Lo, “Ecosystem of Large Language Models for Code,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.16746>
- [10] V. R. B. G. Caldiera and H. D. Rombach, “The Goal Question Metric approach,” *Encyclopedia of software engineering*, pp. 528–532, 1994.
- [11] Hugging Face Inc., “Hugging Face Hub documentation — huggingface.co,” <https://huggingface.co/docs/hub/index>, [Accessed 28-10-2024].
- [12] “Gemini 1.5 Pro — console.cloud.google.com,” <https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-1.5-pro-preview-0409>, [Accessed 28-10-2024].
- [13] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [14] J. Sim and C. C. Wright, “The kappa statistic in reliability studies: use, interpretation, and sample size requirements,” *Physical therapy*, vol. 85, no. 3, pp. 257–268, 2005.
- [15] J. Landis, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, 1977.
- [16] PyTorch Foundation, “PyTorch Hub — pytorch.org,” <https://pytorch.org/hub/>, [Accessed 05-11-2024].
- [17] “Papers with Code - The latest in Machine Learning — paperswithcode.com,” <https://paperswithcode.com>, [Accessed 05-11-2024].
- [18] “TensorFlow Hub — tensorflow.org,” <https://www.tensorflow.org/hub?hl=es>, [Accessed 05-11-2024].