Model Card Metadata Collection from Hugging Face to Foster Multidisciplinary AI Research: A Dataset

Muhammad Asif Suryani[a], Saurav Karmakar[b], Brigitte Mathiak[c] and Philipp Mayr[d]
Knowledge Technologies for the Social Sciences
GESIS – Leibniz-Institut für Sozialwissenschaften, Köln, Germany

Keywords: Hugging Face, Metadata Exploration, Metadata Collection, Large Language Models, Research Data Management, Multidisciplinary Research, Dataset.

Abstract: Metadata features generally exhibit valuable meta information which may facilitate researchers in their tasks. Several studies incorporated scholarly metadata by highlighting its usefulness in certain granularity to assist numerous research tasks. The emergence of Large Language Models (LLMs) has brought an exciting change in the field of Artificial Intelligence (AI) and Machine Learning (ML), which is equally supported by Open Science initiative and FAIR principles. One of the prominent platforms, which ensures the availability of these models to research communities is the *Hugging Face*. It provides democratized access to models while experiencing rapid growth as a repository. As of March 2025, Hugging Face hosts more than 1.4 million models, which were 0.5 million approximately in February 2024. In this dataset paper, we provide information on a large fraction of Hugging Face model cards. Our dataset comprises of a wide range of metadata features which showcase the meta information about each model card. In this work, we aim to provide democratized access to a collection of diverse metadata features from Hugging Face model cards and present an insightful overview of these cards by leveraging the metadata to support the research communities by facilitating model adoption.

1 INTRODUCTION

Open Science is one of the driving factors, which has transformed the scientific disciplines by supporting the accessibility to research artifacts i.e. publications and machine learning models. This initiative is equally supported by the research communities by sharing research artifacts at various platforms to facilitate the information dissemination. In this ecosystem both researchers and repositories are of equal importance as one is going to share the resources and later aims to democratize the access to these resources under certain guidelines to ensure technology is for all (Warzel et al., 2020; Dang et al., 2023).

Researchers across the domains are continuously developing and tailoring intelligent solutions leveraging large language models (LLMs), which has transformed the scientific disciplines recently. These models are actively adopted by the research communities. These models are available at various repositories such as Hugging Face[1] and GitHub[2]. However, considering LLMs Hugging Face is the one of the prominent entity which facilitates the users by hosting state-of-the-art models and provide access to them by its user friendly interface and features. Hugging Face is widely adopted by the research communities due to availability and accessibility of diversified models (Yang et al., 2023; Face, 2025).

These models are becoming essential part of research and development in every domain by their coverages to diverse applications. As Hugging Face asks information about these models during the curation process to collect certain metadata features about these models to enhance the user experience similar to scholarly metadata features. These metadata features are helpful for the users to search for the appropriate model conveniently. These metadata features provide an abstract overview about these models which could facilitate the users in the model adoption process. However, this task is quite challeng-

[a] https://orcid.org/0000-0003-1669-5524
[b] https://orcid.org/0009-0007-0124-5316
[c] https://orcid.org/0000-0003-1793-9615
[d] https://orcid.org/0000-0002-6656-1658

[1] https://huggingface.co/
[2] https://github.com/

Suryani, M. A., Karmakar, S., Mathiak, B., Mayr and P.
Model Card Metadata Collection from Hugging Face to Foster Multidisciplinary AI Research: A Dataset.
DOI: 10.5220/0013571800003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 583-590
ISBN: 978-989-758-758-0; ISSN: 2184-285X

ing and crucial because accessing these models are generally expensive with respect to time. So these metadata features could play an essential role in tackling the rapid increase in the number of models and facilitate the model discovery process concerning diversified research tasks (McMillan-Major et al., 2021; Face, 2025; Yang et al., 2024).

This paper focuses on Hugging Face model cards by presenting coverage of associated metadata features by targeting two timelines between February 2024 to March 2025. Besides, this study also aims to cover insights where such collection of metadata features could be useful and research communities may benefit from it. In addition, it also highlights the dynamics of metadata features to leverage the latest trends prevailing over the Hugging Face including:

- Information on number of model cards over two stages and by which factor model cards are increasing.
- How many models provide complete metadata features?
- How does the adaptability of model cards can be indicated by number of downloads and likes?
- Distribution of model cards by downloads over two stages.
- What are the most influential models in terms of downloads, top k models by month.
- How many models provide information about the licensing?
- How many models provide multilingual support?
- How many models link to external scholarly repositories such as arXiv?

This study is organized as later section presents metadata feature coverage by indicating the set of features followed by related work. The fourth section provides metadata feature collection and by presenting initial explorations. The paper concludes by providing community impact of this study followed by discussions and conclusion.

2 METADATA COVERAGE

Metadata Coverage briefly enlist the heterogeneous metadata features associated with the Hugging Face model cards. These features leverage wide range of technical and generic information and it is important to mention that these features are set by Hugging Face as default but it is not necessary that each model card must carry all of these features. For uniformity and simplicity all the standardized features with default naming conventions are described as available in the data collection (Face, 2025; Suryani et al., 2025):

- The huggingface_link directs to the landing page of model card.
- The modelId/id indicates the name of model and hierarchy which could lead to key references.
- The pipeline_tag presents the primary task associated with each model card.
- The downloads feature is the key aspects which represents the number of times a model has been downloaded, in earlier settings it indicated the cumulative downloads, but currently it reflects only last month's downloads.
- The likes indicates number of likes a model has received over its life time depicted cumulatively.
- The library_name specifies the library or framework associated with each model card.
- The private feature indicates about the access status of the repositories, which is generally set to "False" by default.
- The createdAt: feature shows the timeline but normally it has been observed that it inclined towards date of access.
- The tags features is the key among metadata features as it provides further useful information associated with the model card such as license, dataset, languages, parameters and arXiv ids. But in case no information is provided by default, it carries"region:us" or "region:eu".
- The trendingScore indicates trends regarding model cards introduced recently and its minimum value is 0 and maximum is 999 as observed. The mechanism to calculate it yet not clear.

The information exhibited by these set of features has tendency to provide deeper insights about these models and access to metadata features of model cards seem interesting and may help the research communities to strive for their relevant models. Furthermore, these metadata features cumulatively portray meta information that is essential for supporting model adoption in multidisciplinary research.

3 RELATED WORK

In this problem setting, it is equally essential to highlight the studies where data sources from scholarly publications and model cards features are targeted to facilitate the user experience alongside research data management practices. These metadata features in

general tends to provide an abstract overview and has become an integral part of research ecosystem.

Scholarly metadata is available at numerous repositories which covers certain granularities. Digital Bibliography and Library Project (DBLP)[3] performs metadata acquisition from scholarly resources which covers wide range of features and it is currently one of the largest repositories considering informatics and interdisciplinary research. OpenAlex[4] is also one of the biggest repository which aim to collect and curate scholarly metadata by focusing numerous scientific disciplines. Research Organization Registry (ROR)[5] streamlines metadata features targeting research organization across the globe. These repositories aim to cover distinct set of metadata features and ultimately contribute towards better community experience, discoverability, accessibility and adoption of technological advancement across research domains. Moreover, targeting dataset Zenodo[6] and Kaggle[7] are the prominent repositories which host heterogeneous data sources to facilitate the research communities under FAIR principles thus promoting open science. It is also important to mention that ROR releases its data dump every month (Registry, 2025).

Due to large increase in the number of publications, there has been numerous studies targeting scholarly metadata to assist the researchers by network exploration. The study utilizes Heterogeneous Information Networks (HIN) which has ability to indicate potential relationships such as citation links, author collaborations, and research areas. The populated network is being explored by random walk strategy which simulates the relevance between papers (Du et al., 2020). In addition, another study incorporates a large number of entities such as authors, papers, citations, etc. Such network exploration models diverse data and has the potential to uncover patterns across scholarly resources (Liu et al., 2018).

Recently, an exploratory study targeting Hugging Face was conducted, which considers heterogeneous metadata features of model cards. The study presents an insightful overview of model cards, using downloads and likes as the primary features for segregation. It also presents an overview of linking scholarly repositories and modeling the information into a Heterogeneous Information Network (HIN). As such informative representations highlight the importance of linking scholarly repositories at metadata level while emphasizing on the need of scholarly repository harmonization (Suryani et al., 2024).

Moreover, a study targeting Hugging Face presents how its ecosystem struggles in providing a categorization for the models targeting software development. It discusses an approach which may fill this gap by automating the classification of models for software development tasks, initially extracting relevant information including documentation and relevant tags for these model cards (Di Sipio et al., 2024).

However, another study highlights that lack in transparency across these models can confront issues related to bias, fairness and potential legal risks. It also provides insights against these models by examining model descriptions, datasets for pre-training, training biases and licenses. The results depicts that there has been limited transparency regarding training datasets, biases, and licenses, with several licensing violations and insists on the need for improved transparency (Pepe et al., 2024; Castaño et al., 2023).

In a technical report on 4chan which gather observations regarding various sensitive topics. This report provide the approach being adopted to collect the data from its various boards and briefly discuss how such dataset can drive the social studies revolving around digital behavior (Culbert, 2023).

4 MODEL CARD DATASET COLLECTION AND EXPLORATION

Dataset are one the key elements in driving the research specifically in this era of Large Language Models (LLMs). Dataset are generally made available across the open repositories. However, it is equally important to make these datasets discoverable and accessible, in addition to providing useful meta information, such as a READMEs. The recent study on data searching indicates that it is still difficult to search for appropriate dataset (Hulsebos et al., 2024).

4.1 Metadata Collection and Timeline

Over the course of this study, we aim to collect metadata features of the Hugging Face model cards available at Hugging Face via Hugging Face API. The brief overview is presented in Figure 1 which commence with the metadata retrieval and extracts unstructured metadata features from the Hugging Face API which need to be further processed to convert into structured format such as CSVs. It ensures that all the features remain intact with their respective descriptions as presented in section 2.

[3]https://dblp.org/
[4]https://openalex.org/
[5]https://ror.org/
[6]https://zenodo.org/
[7]https://kaggle.com/

Information parser handles all the inconsistencies by following pre-processing steps which covers handling encoding issues, missing values, parsing nested information and remove redundancies to make sure that data is consistent. Moreover, these processes are responsible of providing metadata features in a structured format because metadata modeling is the key process which thoroughly study all the feature set and tends to grasp useful information from metadata features individually and collectively by forming informative links among features. For example, downloads and likes values depict model adaptability. Dataset exploration module provides the interesting insights covering distribution of models by features, top model cards by timeline and pipeline/library tags which has ability to drive the interest of research communities by facilitating AI adoption across the domains.

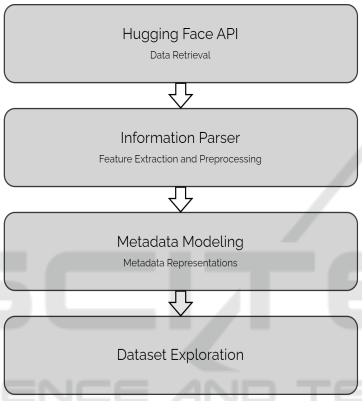

Figure 1: Model Card Metadata Collection Overview.

The metadata acquisition process commenced in February 2024 by collecting the metadata features and later it has been divided into two stages by timeline i.e. February 2024 to April 2024 and September 2024 to March 2025. The selection of these timelines is based on the availability and interpretability of the model cards metadata features. For example, Hugging Face currently provides month-wise downloads for model cards, whereas previously it was providing cumulative downloads. Over these timelines we aim to present a consolidated insights to research communities that how metadata features could be helpful in the model discovery process. Moreover, to support the study a couple of dataset instances are available at Zenodo (Suryani et al., 2025).

4.2 Dataset Exploration

Data Exploration section presents the comprehensive insights about the Hugging Face model cards by exploiting the metadata features over two timestamps.

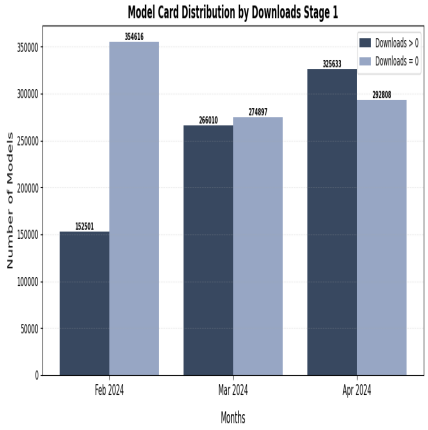

Figure 2: Model Card Overview by Downloads Stage 1.

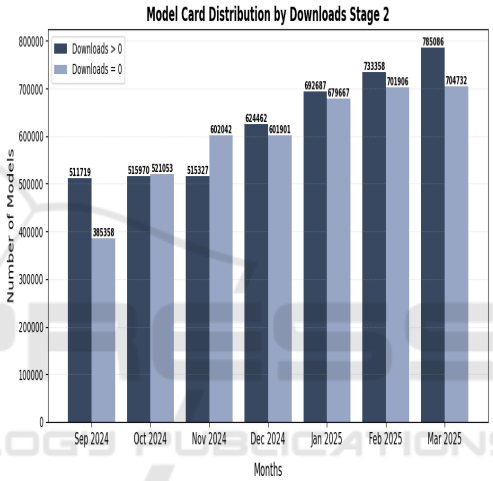

Figure 3: Model Card Overview by Downloads Stage 2.

The Figure 2 and Figure 3 present the distribution of model cards, in which model cards are divided into two groups such as models with zero download and models with downloads greater than zero.

It is evident that almost half of the model cards are having no downloads which is an important aspect for the research communities as well as for the repository.

Moreover, Figure 4 and Figure 5 demonstrate the distribution of model cards considering likes over the two stages respectively. Similar to downloads the model cards are segregated into two sets i.e. models with likes and models with no like respectively. These insights showcase the adoption of this feature by the communities and hence cannot be used as a primary parameter to measure the impact of model cards. However from figures one can realize a substantial increase in total number of models over time. Hence, it is important to mention that a considerable amount of models having zero downloads which may indicate their adaptability across research communities. So the focus will be on the model cards having

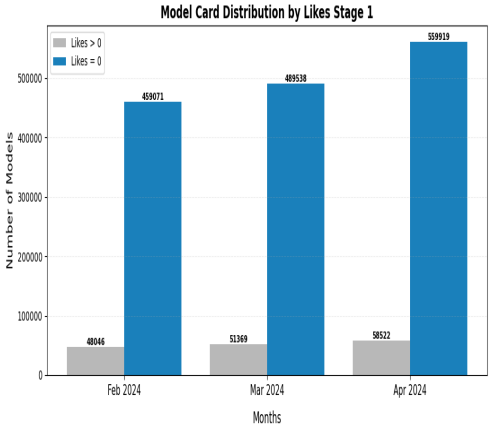

Figure 4: Model Card Overview by Likes Stage 1.

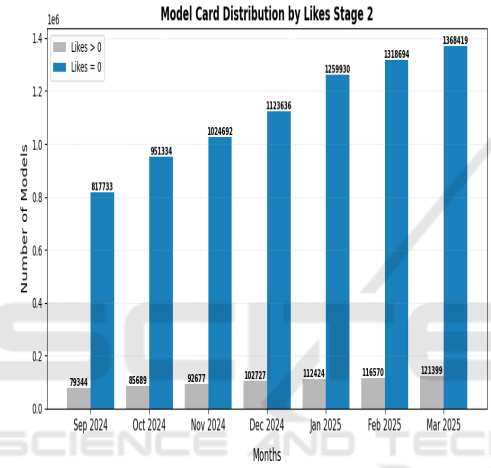

Figure 5: Model Card Overview by Likes Stage 2.

downloads greater than zero.

Besides, downloads and likes, tags is also an important feature which comprises of nested information presenting useful overview about the model cards. For example arXiv ids, dataset mentions, license information, language support and many more. But it is also essential to mention such information is not uniform across the model cards and depends upon the curation process.

Hence, Figure 6 and 7 effectively demonstrate the total number of model cards over these timestamps and also presents the available auxiliary information across these models.

These figures highlight how a relatively small number of model cards are providing useful information, it can be seen that almost one third or even lower number of model cards are providing information about the licenses which is a crucial indicator in the adaptability process. Moreover, it is also important to highlight that a very low number of models are providing information about the relevant datasets and

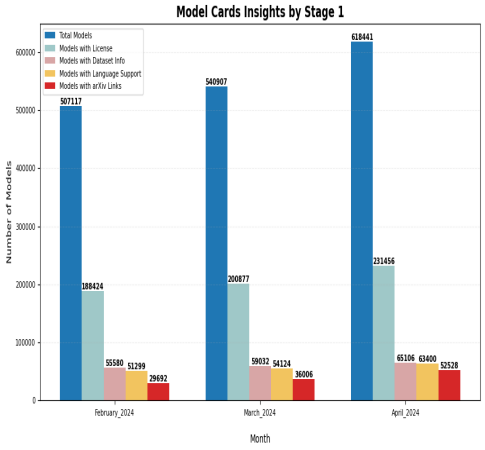

Figure 6: Model Card Overview by Tags Features Stage 1.

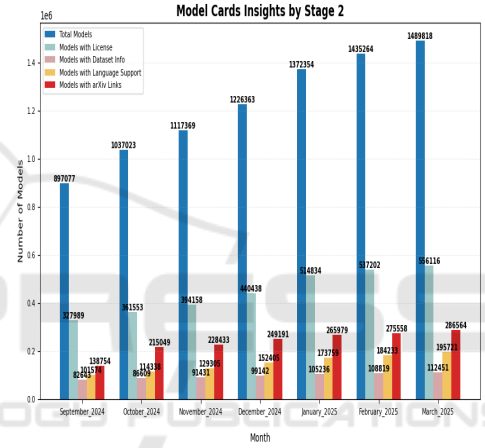

Figure 7: Model Card Overview by Tags Features Stage 2.

that trends remains persistent when it comes to models mentioning publications. So, availability of such valuable meta information will defiantly enhance the overall reproducibility and user experience.

Furthermore, it is also important to target features covering technological aspects concerning model cards, as access to such information holds key importance in the adoption process. The Figure 8 and Figure 9 present interesting insights targeting library and pipeline features.

The aim of this representation is to indicate the trends of library and pipelines which are being exploited by the research communities.

Moreover, for simplicity we present the top 10 most frequent items. The trends across both timestamps remain similar, with “transformers” and “text-generation” being the most widely adopted under library and pipeline features respectively. But in second stage there is fluctuation is observed considering image based pipelines. Hence, from these exploration it is evident that these metadata features are capable

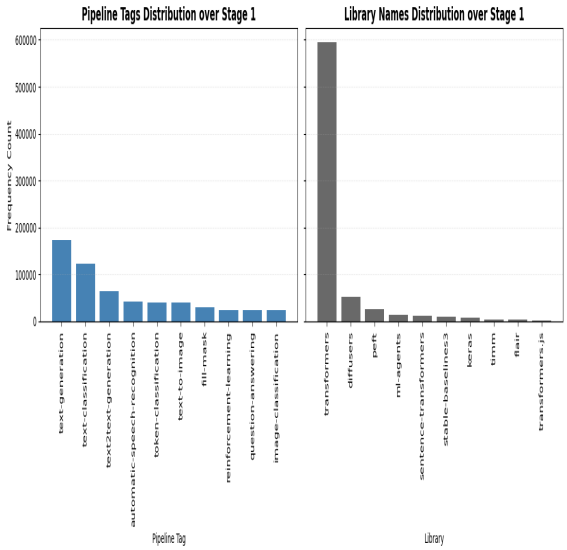

Figure 8: Pipeline and Library Overview Stage 1.

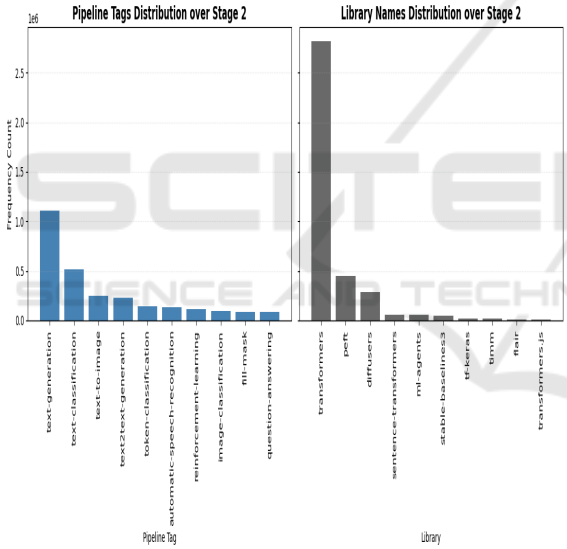

Figure 9: Pipeline and Library Overview Stage 2.

of providing useful insights towards research communities which could pave the way for the indulgence of these model across multidisciplinary research while assisting in the model discovery process.

5 COMMUNITY IMPACT

It is essential to discuss how model cards metadata exploration can provide valuable insights by targeting pipelines, libraries, language coverage, and datasets. These trends may reflect model popularity, preferred licensing, and dataset diversity, enabling researchers to align their work with the evolving needs of the community. Furthermore, tracking model cards with arXiv ids can help in identifying models with papers and foster research collaborations by enhancing reproducibility. These insights also support data driven decision making and facilitate future research and open science initiatives within the research communities.

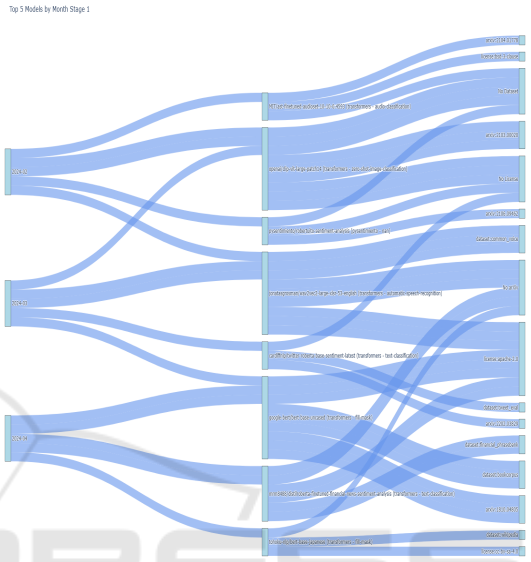

Figure 10: Top 5 Model Cards in Stage 1.

To support these arguments, Figure 10 presents the top five model cards from stage 1 which illustrates a shift in technology as well as the adoption of various model cards across research communities. Moreover, it also highlights the potential contributors driving advancements in research and technology.

Similarly, Figure 11 also illustrates the top models by each month across stage 2 by revealing notable informative trends in model adoption, pipeline preferences, and dataset utilization. In stage 1, a strong presence of transformers based models has been observed, particularly targeting text-classification and text-generation tasks. The presence of models with paper in stage 1, although not dominant, suggests that some models originate from published research. However, the limited diversity in datasets and licenses can be crucial for the community driven innovation.

However, stage 2 showcases a broader shift in model diversity and complexity. Notably, speech recognition and multi-modal models emerge alongside traditional text-based models, signifying growing interest in speech and vision applications. The adoption of models utilizing diffusers for generative tasks is also notable, indicating a rise in image generation and diffusion-based techniques. Furthermore,

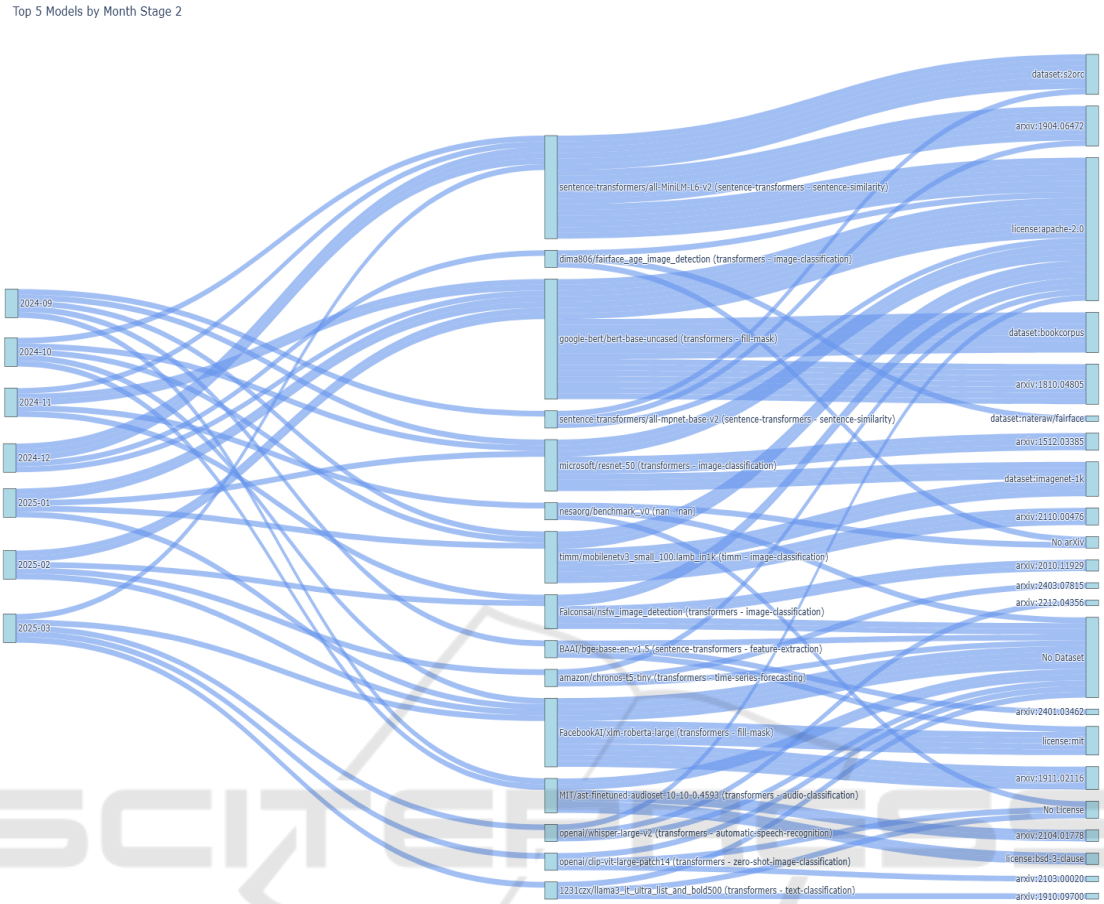

Figure 11: Top 5 Model Cards in Stage 2.

the higher number of models with associated publications in this stage suggests that research driven contributions are becoming more prevalent and may foster stronger linkages between academia and industry.

Moreover, this exploration highlights the importance of metadata features within model cards and emphasizes the need to collect such metadata as a dataset to support the research community in the model adoption process. These insights hold significant implications for both the research communities and the Hugging Face ecosystem. By analyzing these trends, this exploration may serve as a valuable reference point for researchers across various domains, offering guidance in shaping future directions and addressing the needs of model development aligned with open science initiatives and research data management.

6 DISCUSSION AND CONCLUSION

In this era, data driven applications are getting indispensable across research communities, elevating the need for studies that focus not only on data acquisition but also on necessary exploration. Such efforts not only provide insightful overview but also help in managing rapidly evolving technological dynamics. These data sources can play a crucial role in driving and reshaping research practices.

This study aims to democratize access to the metadata features of Hugging Face model cards, offering as a valuable data source in various meaningful configurations. Because such exploration alongside Hugging Face existing searching mechanism will be ideal for researchers to search for the appropriate models conveniently. However, it is also important for Hugging Face to come up with some deprecation policy to further optimize the user experience related to model cards. In future more exciting dataset instances will

be shared on Zenodo to facilitate the adoption of AI across research communities.

This work have exciting future prospects such as Model Recommender System, Hugging Face Model card Leader-boards which provide users valuable recommendations based on metadata features of model cards alongside technological description available on the landing pages of each model cards. Moreover, inclusion of current state of development of model cards in the recommendations will be interesting and indicate the current state of development of the models.

Furthermore, an exciting direction is the harmonization of research artifacts across scholarly repositories, which would enrich the research ecosystem with more linked information for researchers such as "Authors with Models" and "Organization with Models'. In addition, exploiting model card provenance will also be interesting and may yield valuable insights.

ACKNOWLEDGEMENTS

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), NFDI4DS (Grant number 460234259). Authors also acknowledge the Hugging Face as data sources and also thanks the individuals involved in this research.

REFERENCES

Castaño, J., Martínez-Fernández, S., Franch, X., and Bogner, J. (2023). Exploring the carbon footprint of hugging face's ml models: A repository mining study. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12. IEEE.

Culbert, J. H. (2023). 4tct, a 4chan text collection tool. *arXiv preprint arXiv:2307.03556*.

Dang, V.-N., Aussenac-Gilles, N., Megdiche, I., and Ravat, F. (2023). Interoperability of open science metadata: What about the reality? In *International Conference on Research Challenges in Information Science*, pages 467–482. Springer.

Di Sipio, C., Rubei, R., Di Rocco, J., Di Ruscio, D., and Nguyen, P. T. (2024). Automated categorization of pre-trained models for software engineering: A case study with a hugging face dataset. *arXiv preprint arXiv:2405.13185*.

Du, N., Guo, J., Wu, C. Q., Hou, A., Zhao, Z., and Gan, D. (2020). Recommendation of academic papers based on heterogeneous information networks. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE.

Face, H. (2025). Hugging face apis. https://huggingface.co. Accessed: 2025-02-20.

Hulsebos, M., Lin, W., Shankar, S., and Parameswaran, A. (2024). It took longer than i was expecting: Why is dataset search still so hard? In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics*, pages 1–4.

Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., and Xia, F. (2018). A survey of scholarly data visualization. *Ieee Access*, 6:19205–19221.

McMillan-Major, A., Osei, S., Rodriguez, J. D., Ammanamanchi, P. S., Gehrmann, S., and Jernite, Y. (2021). Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the huggingface and gem data and model cards. *arXiv preprint arXiv:2108.07374*.

Pepe, F., Nardone, V., Mastropaolo, A., Bavota, G., Canfora, G., and Di Penta, M. (2024). How do hugging face models document datasets, bias, and licenses? an empirical study. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension*, pages 370–381.

Registry, R. O. (2025). Ror data.

Suryani, M. A., Karmakar, S., and Mathiak, B. (2024). Exploration of hugging face models by heterogeneous information network and linking across scholarly repositories. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 371–386. Springer.

Suryani, M. A., Karmakar, S., Mathiak, B., Mutschke, P., and Mayr, P. (2025). Hugging face model cards metadata dataset.

Warzel, D., Fitzmartin, R., Zhou, F., et al. (2020). Fair data sharing: the roles of common data elements and harmonization. *Journal of biomedical informatics*, 107:103421.

Yang, Z., Shi, J., Devanbu, P., and Lo, D. (2024). Ecosystem of large language models for code. *arXiv preprint arXiv:2405.16746*.

Yang, Z., Wang, C., Shi, J., Hoang, T., Kochhar, P., Lu, Q., Xing, Z., and Lo, D. (2023). What do users ask in open-source ai repositories? an empirical study of github issues. In *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, pages 79–91. IEEE.