Exploration of Hugging Face Models by Heterogeneous Information Network and linking across Scholarly Repositories

Muhammad Asif Suryani[1][0000−0003−1669−5524], Saurav Karmakar[1][0009−0007−0124−5316], and Brigitte Mathiak[1][0000−0003−1793−9615]

Knowledge Technologies for the Social Sciences
GESIS- Leibniz-Institut für Sozialwissenschaften, Köln, Germany
{asif.suryani,saurav.karmakar,brigitte.mathiak}@gesis.org

Abstract. With the pervasive integration of Machine Learning (ML) focusing complex tasks across various domains, generally respective models and datasets are made available in numerous scientific repositories such as Hugging Face, GitHub for recognition and understanding within the research communities hence supporting open science initiative. However, the adaptability of these repositories is increasing among users hence raising a concern about the usability of these models and datasets effectively. Therefore, it is necessary to explore these repositories and compile comprehensive information that could facilitate users as well as repositories itself. Hugging Face, a leading repository, aims to furnish a platform that organizes and presents detailed information on models and datasets employed in research. As its adoption escalates within the research communities, the necessity to delve into such repositories becomes crucial to offer researchers valuable insights and promote efficient knowledge dissemination. This study focuses on exploring Hugging Face, particularly its machine learning models by exploiting various relevant features and the insights are presented in a Heterogeneous Information Network (HIN). Our research not only demonstrates the effectiveness of the available models on Hugging Face but also highlights potential links to relevant scholarly repositories by highlighting the significance of exploration which could contribute to the future integration of repositories, facilitating a more unified and accessible framework for scientific research.

Keywords: Data Exploration · Hugging Face · Machine Learning Models · Heterogeneous Information Network · Repository Exploration

1 Introduction

The exponential growth in Artificial Intelligence (AI) and Machine Learning (ML) research has significantly impacted the scientific domains, leading to an unprecedented increase in computing-related scientific publications and nourish the interdisciplinary research paradigms. These publications not only contribute in acquiring knowledge but also introduce innovative models and datasets which could be crucial for future research directions in respective domains. As metadata within these publications serves as the critical feature for cataloging, searching, and referencing these contributions, enabling researchers to look for their relevant information efficiently [9, 10].

However, these resources are effectively organized and made available by various repositories to facilitate the research communities. There have been numerous repositories each serving distinct roles within the research ecosystem. Kaggle[1], renowned for hosting competitions and datasets, provides a platform for practical data science and ML challenges. GitLab[2], on the other hand, offers a more comprehensive code repository service that facilitates version control and collaborative project management, including the sharing of AI models and datasets. Hugging Face[3] stands out by specifically targeting the AI research community, offering a specialized repository for AI models and datasets. As these repositories are playing a significant role in providing a state of the art platform to disseminate the models and datasets conveniently with the research communities which generally seems beneficial for the models itself and also for the communities by supporting the open science initiative [3].

Hugging Face has rapidly gained prominence within the AI research community, attributed to its user-friendly interface and comprehensive collection of models and datasets. Unlike traditional repositories, Hugging Face focuses on democratizing access to state-of-the-art AI technologies, encouraging open-source contributions, and facilitating easy implementation and bench-marking of models. Its significance extends beyond a mere repository and it acts as a hub for collaboration, knowledge exchange, and innovation in respective domains. Hugging Face aims to effectively comprehend their models by prompting curators for certain features which could be beneficial in categorizing the models, besides maintaining certain informative features. So collectively these diverse features are made available against each of the model card [2].

These features could provide essential information such as number of downloads, likes, library names, pipeline, citation and tags. Exploration of these potential features for model cards could be interesting and may bring insights by highlighting potential research trends to bridge the repositories and support repositories harmonization. However, to explore these models cards and their potential relationships, a diverse data model and heterogeneous representation is essential. Heterogeneous Information Network (HIN) has been used in various studies modelling diverse features and their relationships. In this problem setting, complex interconnections between features of model cards and respective links to publications can be effectively represented and explored through network representations of repositories such as Hugging Face and arXiv. The network enable a novel approach to navigate the large repository landscape, hence uncovering hidden patterns, and facilitating targeted searches. As sample studies utilizing HIN for repository exploration have demonstrated its potential

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

in enhancing and understanding the data relationships, contributing to more efficient and insightful research methodologies. [7, 12].

However, exploring relevant features of Hugging Face models could bring more descriptive insights at the repository level, which defiantly will be beneficial for the communities and enhance the overall user experience. So to address this challenging problem, it is necessary to consider wide range of diverse available features from these scientific models to further broaden the information extraction spectrum. As it will bring desired information alongside respective contextual information, which could be modelled into an information network. In this paper, we conducted a study to explore machine learning research model available in Hugging Face, which aims to provide insights to users by modelling diverse features into a Heterogeneous Information Network. Primarily, here we address certain research questions, including but not limited to:

- How can descriptive statistics of Hugging Face models (downloads, likes) reveal trends in the adoption and usefulness of machine learning models?
- What are the most influential models and libraries in Hugging Face, and whether it contributes toward cross repository linkages.
- How does representing the model information of Hugging Face repository as a Heterogeneous Information Network (HIN) facilitate the exploration of relationships between models, and research articles?
- What insights can be acquired by analyzing the network structure, and how does this structure reflects the contributions of researchers to open science initiatives?
- Does any collaborations between different research entities reflect and flourish the interdisciplinary research by model citations?
- What role does metadata (e.g., model information, tasks, tags) plays in enhancing the usability and discoverability of models within the Hugging Face?

This study is organized as: following section will present the related work in this problem setting. The third section presents the exploratory study by providing the data model for heterogeneous information network. Later sections provide discussion on the collected insights followed by conclusion and future directions.

2 Related Work

Over the course of interdisciplinary research, the involvement of machine learning models has showcased substantial adaptation. Such initiatives are predominantly facilitated by open-source repositories. These repositories provide a platform for researchers to understand and utilize machine learning models in their research. Numerous research studies aim to highlight the usefulness of these repositories. In addition, effective relevant repositories exploration may contribute towards enhanced user experience.

Open-Source Software (OSS) projects are widely available on various social coding platforms such as GitHub and GitLab. However, with the emergence of machine learning models for interdisciplinary research tasks, specialized platforms like Hugging Face Hub (HFH) have become essential by focusing on sharing datasets, pre-trained models, and applications. HFH is rapidly growing, currently maintaining more than 400K repositories accessible via an API. Besides the provided API, there is a need for solutions to facilitate data collection and the exploration of HFH's different facets. To address these tasks, researchers introduced HFCommunity, an extraction process for HFH data and a relational database designed to facilitate the empirical analysis of the growing number of ML projects [1].

Hugging Face (HF) has emerged as a central repository for sharing and developing machine learning and artificial intelligence models. This study provides insights into various aspects of HF focusing on the carbon emissions and ML model maintenance. It aims to guide researchers in mining software repositories within the HF ecosystem, offering a framework that promotes the responsible and sustainable advancement of ML. Additionally, the study also fosters a deeper understanding of the broader implications of ML models [2].

The development of natural language processing tools involves various aspects, including datasets, models, individual skills, and the crucial task of following guidelines for documentation for potential future use. The adoption of standards for the documentation process enhances accessibility across the research community. This study presents two case studies to showcase reusable documentation templates on Hugging Face. It describes the key processes of developing templates, focusing on relevant stakeholders and entities [9].

Recent studies describe how transformer architectures have significantly empowered natural language processing. Additionally, enhanced pre-training processes have effectively broaden the applications of these models to address diverse scientific tasks across research domains. Furthermore, the study also highlights the usefulness of transformers and underscore the value of research repositories, emphasizing the impact of these advancements on scientific communities [14].

A recent study focused on the usability of pre-trained models (PTMs), specifically within the Hugging Face ecosystem. The study presents an empirical investigation of PTM reuse by interviewing twelve practitioners from Hugging Face to understand best practices and challenges in PTM re-usability. Additionally, the study enables researchers to model the decision-making process by identifying key attributes that contribute to the re-usability of models in the Hugging Face ecosystem [6].

A recent study analyzed the 1,417 hugging face machine learning models and their respective dataset for carbon footprint. The study aimed to provide insights and recommendations to optimize the carbon efficiency of these models. Focusing on the Hugging Face Hub API, the study addressed two research questions regarding the measurement and reporting of carbon emissions. The findings indicate a slight decrease in reported carbon footprint over the past two years and reveal correlations between carbon emissions and attributes such as model size, dataset size, ML application domains, and performance metrics [4]

The number of research papers across various domains is increasing due to advancements in science and technology, making it challenging for researchers to find relevant literature. This study proposes a novel approach based on Heterogeneous Information Networks (HIN) that targets potential relationships such as citation links, author collaborations, and research areas. By constructing network and using a random walk strategy to simulate natural sentences, the approach effectively discovers relevance between papers [5]. With the rapid growth of digital publishing, efficiently visualizing scholarly data has become increasingly demanding. This data includes millions of raw data points such as authors, papers, citations, and scholarly networks. Various visualization techniques can be applied to better represent data structures and uncover hidden patterns. The study introduces the basic concepts and collection methods for scholarly data and provides a comprehensive overview of related visualization tools and techniques [8].

Recently, many works have incorporated auxiliary information into recommender systems to address data sparsity. Heterogeneous information network (HIN) based recommender systems provide a unified approach to integrate various auxiliary information, enhancing the performance and Interpretability. This study delves into the network-based recommender systems, covering concepts, methods, applications, and resources. Finally, the study discusses the potential research directions incorporating Heterogeneous Information Network [7].

3 Proposed Exploration

This sections describes the overall exploration approach being adopted in this study. The Fig. 1 showcase the block diagram of tasks by highlighting the top-down approach being followed for the Hugging Face exploration.

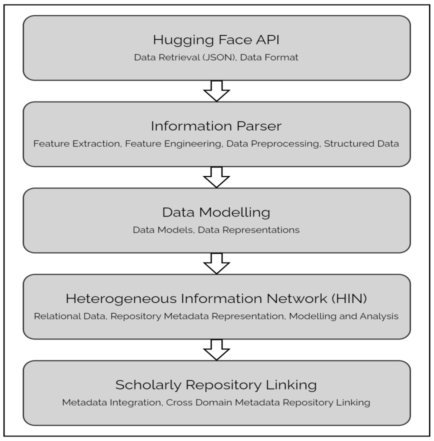
Hugging Face API
Data Retrieval (JSON), Data Format
Information Parser
Feature Extraction, Feature Engineering, Data Preprocessing, Structured Data
Data Modelling
Data Models, Data Representations
Heterogeneous Information Network (HIN)
Relational Data, Repository Metadata Representation, Modelling and Analysis
Scholarly Repository Linking
Metadata Integration, Cross Domain Metadata Repository Linking

Fig. 1. Block Diagram of Hugging Face Exploration Pipeline

The exploration commence with data retrieval by the use of Hugging Face API, targeting models to fetch unstructured data in JSON format. This extracted data includes numerous available metadata features for each model, such as architecture, pipelines, likes, downloads, tags and more. The initial processing ensures that the data is structured and ready for subsequent tasks.

The acquired data is further passed to the information parsing module, which segregates and extracts the relevant feature set for each of the Hugging Face model. However, the feature engineering is applied to transform these features into a meaningful set of interesting features, omitting irrelevant features such as URLs and local IDs. Furthermore, pre-processing techniques are applied to handle missing values, encoding issues, and remove redundant information, ensuring the data is clean, consistent, and ready for further analysis. It has been observed that some models have multilingual tags and duplicate entries, which were removed.

By extensively studying the diverse feature set, we elaborate the diversified data models being proposed for the insightful exploration of Hugging Face machine learning models as shown in Fig. 2 and Fig. 3 respectively. These data models exploit the feature set for each model to comprehensively explore and group the relevant models. These data models are designed to support the proposed research questions and capture the complexities and relationships within the feature set of Hugging Face models.

The first data model, as shown in the Fig. 2, focuses on the relational aspects of models by highlighting connections between models and their relevant features such as models being linked to specific pipeline and libraries respectively. In addition it also indicate the connections between models and their available tags, downloads, likes and citation tag available for each model.

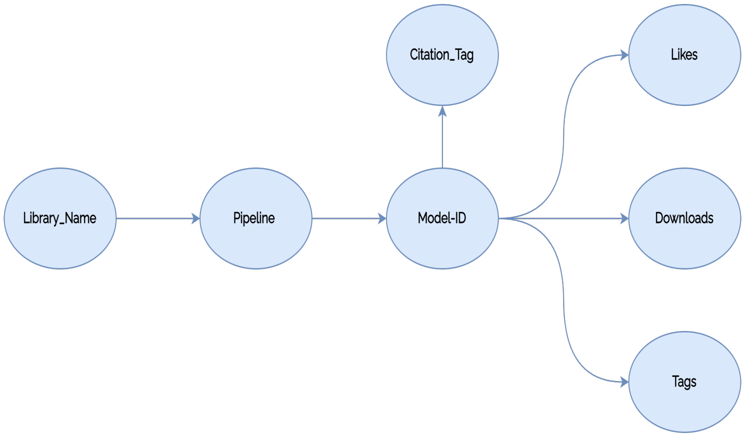

Fig. 2. Data Model for Hugging Face Exploration

The second data model primarily aim to support the cross-domain repository linking between Hugging Face and metadata repository such as arXiv. The feature representations help in understanding the potential patterns and enhance the interpretability and effectiveness of Hugging Face models across metadata repository. The models having a citation tags are considered at this stage for exploration. Beside, scholarly metadata features available against each model, important features such as library and pipelines are also incorporated to explore the cross disciplinary repository linking effectively.

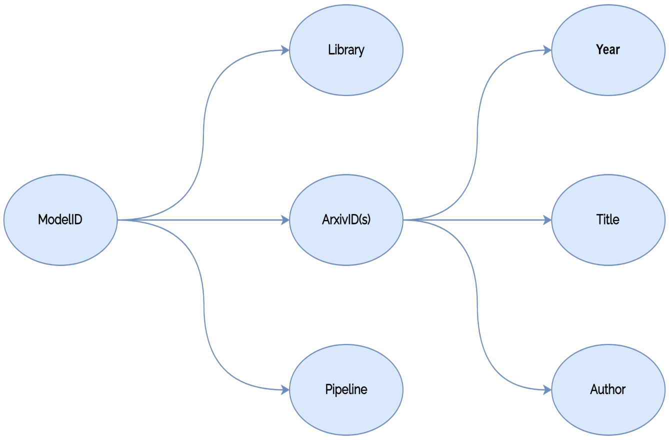

Fig. 3. Data Mode for Repository Linking having Citation Tag

Heterogeneous Information Networks (HIN) are advanced models used to represent and analyze complex data comprising multiple types of entities and relationships. Unlike traditional homogeneous networks that consist of a single type of node and edge, HINs incorporate diverse node and edge types, capturing the multifaceted nature of diverse data. HIN can be defined as a graph $G = (V, E)$ where V denotes the set of nodes (or vertices) and E represents the set of edges.

This diversity allows for a comprehensive representation and understanding of interconnected data, making HINs particularly useful in domains such as social network analysis, bioinformatics, and recommendation systems. The ability to model complex interactions and dependencies within the data provides a powerful tool for addressing intricate research questions, enabling more precise and insightful analyses [7]. So, HINs provide a powerful framework for modeling and analyzing complex relationships. In this problem setting, HINs are constructed by leveraging the proposed data models, which integrate the relational data and metadata of the models and perform comprehensive analyses to explore the structure and dynamics of the models.

Finally, the HINs are used to explore the Hugging Face models within the repository, additionally highlighting potential insights and patterns. Moreover, the second HIN provides the potential linking to external scholarly repositories, enabling metadata integration and cross-domain analysis. In general HIN explorations allows for seamless linking of data across repositories, enhancing the richness and utility of the information. This broadens the scope of analysis and facilitates comprehensive insights by leveraging data from diverse sources, leading to repository harmonization.

4 Results and Discussions

This section offers the comprehensive insight being collected over the course of this study by providing the detailed exploration of Hugging Face models for the selected feature set considering proposed data models.

4.1 Data Description

Hugging Face is the one of the biggest repository which facilitate the hosting and sharing of machine learning models to support the scientific communities. Over the course of this study we collected the metadata concerning hosted models of Hugging Face. The data considered from the exploration is till April 2024 and the dataset is being collected from the Hugging Face API. The dataset in this regard comprise of various potential features. But in this study we focused on *Model-IDs*, *Pipeline tag*, *Library Name*, *Downloads*, *Likes* and *tags*. Moreover, *tags* features carry a wide range of textual keywords attached to the models. These tags usually carry information about potential citation tags, language supports, pipeline, licences and many more and the study aim to consider wide range of these available tags as potential features.

4.2 Descriptive Exploration of Models

The potential feature set for the exploration of Hugging Face are quite diverse. We presents the segregation of Hugging Face models by two quantitative features such as downloads and likes. The Fig. 4 presents the distribution of downloads of hugging face models which clearly states the drops in number of models as the number of download increases. Likewise the Fig. 5 showcases the distribution of likes of hugging face models, which clearly states the huge drops in the number of hugging face models as number of like increases.

However, to further explore the downloads and likes features, we created various ranges for each of the feature and plot the findings in Fig. 4 and Fig. 5 respectively. The Fig. 4 shows the distribution of models for the various downloads ranges. It is evident that a substantial number of models lies in downloads range from 0 to 50 and further decreases as the value of ranges increases. Moreover, Fig. 5 also exhibits similar trends and indicate that a large number of models having likes from 0 to 50 and only one model have more than 10000 likes.

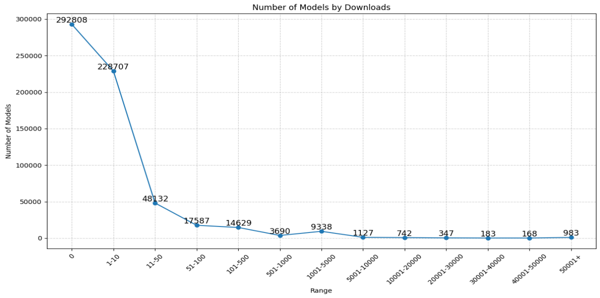

Fig. 4. Downloads Distribution of Hugging Face Models

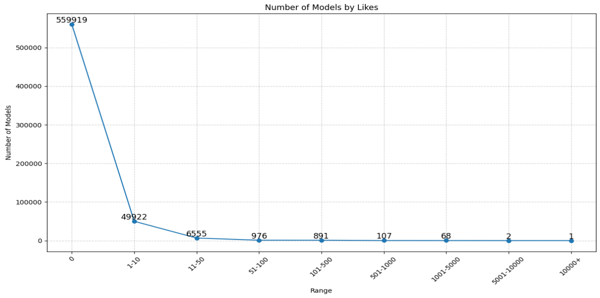

Fig. 5. Likes Distribution of Hugging Face Models

In addition for deeper insights, the most downloaded and liked models also corresponds to scholarly metadata available in the tags features. Thus it further emphasise to broaden the scope of exploration to have more interesting and useful insights. Furthermore, the information is also organized considering downloads and likes and presented in Table 1.

Table 1. Hugging-Face model Segregation

Segregation Criteria	**Model Count**
Downloads > 0	325633
Downloads = 0	292808
Likes > 0	58522
Likes = 0	559919
Both Downloads and Likes > 0	44822
Both Downloads and Likes = 0	279108

The overall findings demonstrate that downloads and likes tends to be the features which cannot be considered solely to explore the Hugging Face models. But could be important features. Hence indicating that for comprehensive exploration it is necessary to consider more features. To support that we collected the tags considering both downloads and likes and processed them against the defined ranges and created word-clouds to anticipate the usefulness of modelling features beyond downloads and likes. The word-clouds for downloads and likes are shown in Fig. 6 and Fig. 7 respectively.

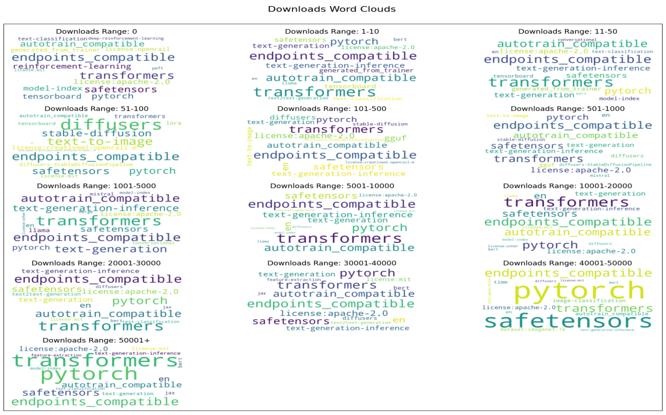

Fig. 6. Word Cloud of Tags for Downloads Ranges of Hugging Face Models

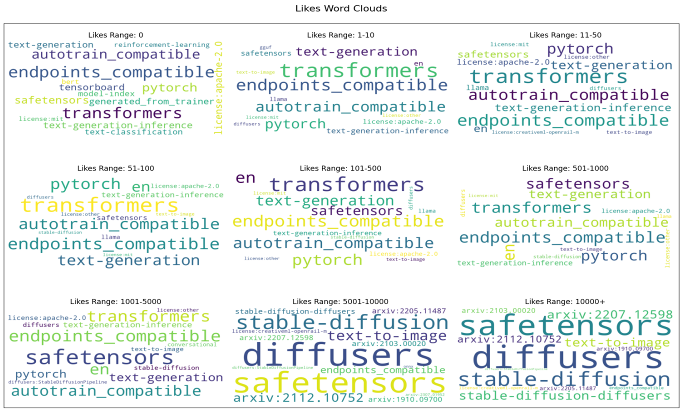

Fig. 7. Word Cloud of Tags for Likes Ranges of Hugging Face Models

As these representations, showcase the need of further exploration to have a deep insight about the models. As in both word clouds the frequent words are names of the pipelines or libraries and also the languages being supported by the model. In addition, there is also an considerable representation of citation tags which could lead to scholarly metadata repository. Finally, this exploration also supports the network modelling direction by keeping the downloads and likes ranges.

4.3 Heterogeneous Information Network Exploration

To further explore the feature set of Hugging Face models in Heterogeneous Information Network (HIN), the previously applied ranges for downloads and likes are taken into account alongside proposed data models. Hence, for simplicity we only showcase the one with best presentable network each for downloads and likes ranges. The criteria adopted for the presentation of the networks are the visibility of diverse nodes and their connections. Here to further enhance the interoperability of network, each type of nodes are color-coded.

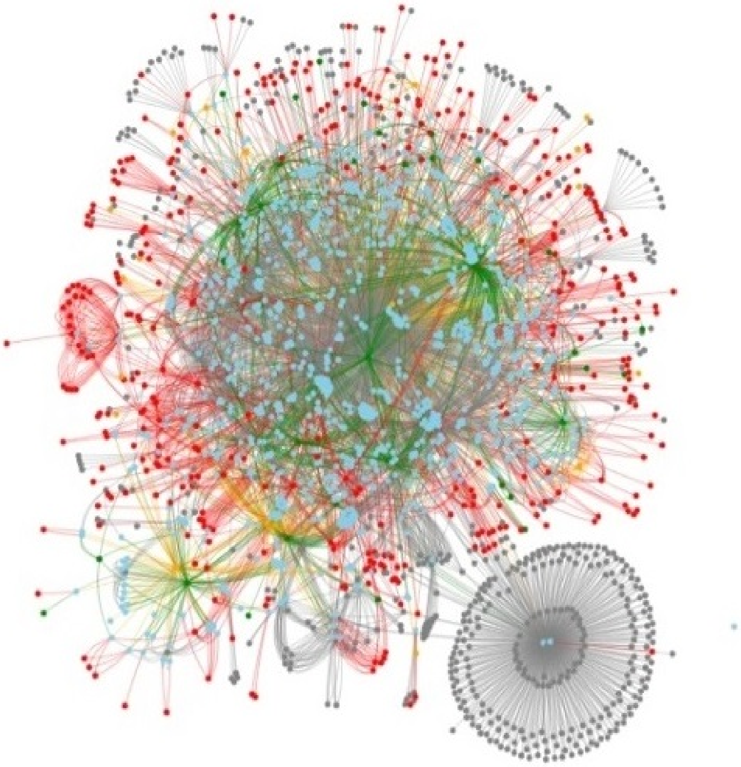

Fig. 8. Network of Download Range(50001 and Plus) using First Data Model

The Fig. 8 showcased the information network presenting the diverse nodes and their interconnections as in the Fig. 8 the red colored nodes indicate the citation tags are available for the models and light blue colored nodes indicate the models and grey nodes are the generic tags for each model. The pipeline nodes and library nodes are represented in green and yellow respectively. For this specific network we picked the last download range i.e. 50001 and plus as it carries 983 models. The number of red colored nodes showcase the potential links of machine learning models to publications over specific download range.

Moreover, a network was also populated following the similar criteria for likes ranges and the exploration is quite evident that there exist a potential exploratory link to scholarly repository as shown in Fig. 9. For the likes, we picked the similar range picked earlier i.e. 501 to 1000 which carries 107 models.

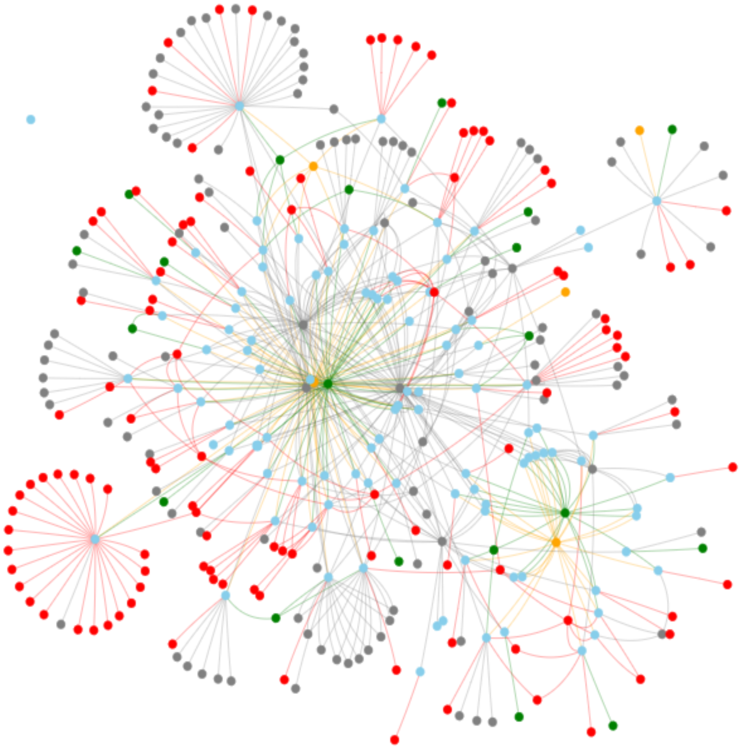

Fig. 9. Network of Likes Range (501-1000) using First Data Model

In addition to support the argument of linking information to scholarly repositories we processed tags features of models to have arXiv ids, we used the Python wrapper for arXiv[4] to collect the metadata. In addition, to further emphasize

[4] https://github.com/lukasschwab/arxiv.py

on the linking between scholarly repositories by considering second data model, we created the networks for both downloads and likes ranges as adopted earlier.

The Fig. 10 and Fig. 11 illustrates information networks by incorporating the metadata from arXiv and Hugging Face model features exploiting data model from Fig. 3. Each node in the network represents a different entity and corresponding link to other types of nodes. The red colored nodes denote arXiv ids, green nodes represent Hugging Face model ids, and blue nodes are the authors and publication dates. However, the connections between these nodes indicate potential relationships, such as authorship or the presence of an arXiv paper id within tags of a model.

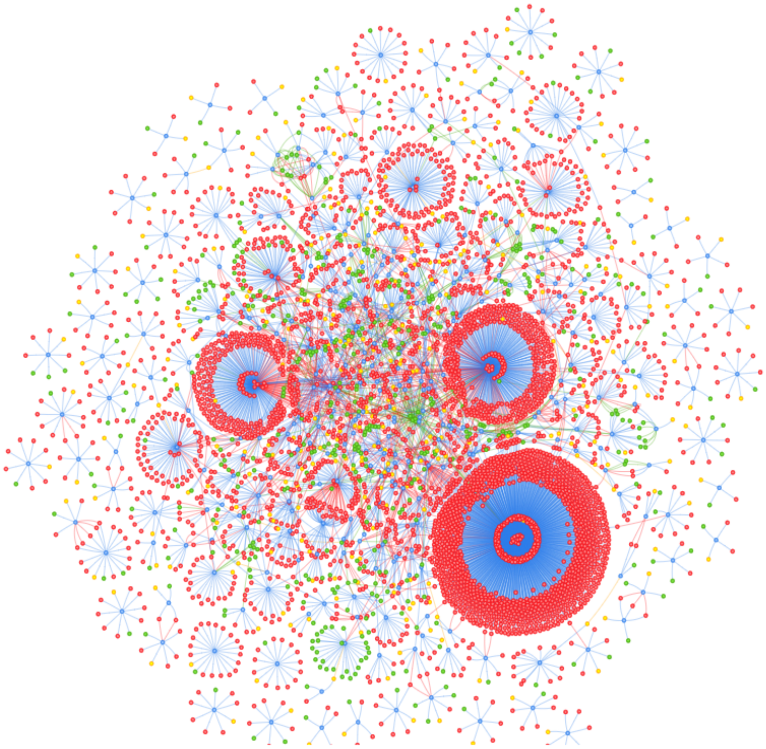

Fig. 10. Network of Downloads range (50001 and plus) models with arXiv tags

These networks demonstrate the direct and indirect inter-connectivity of research publications with respective machine learning models, highlighting how specific papers influence and contribute to the development of these models. In these network representations, the largest portion represents Gemini multimodel publication [13] and second largest representation is from the large language model article [11] which are connected to numerous Hugging Face models considering the picked ranges.

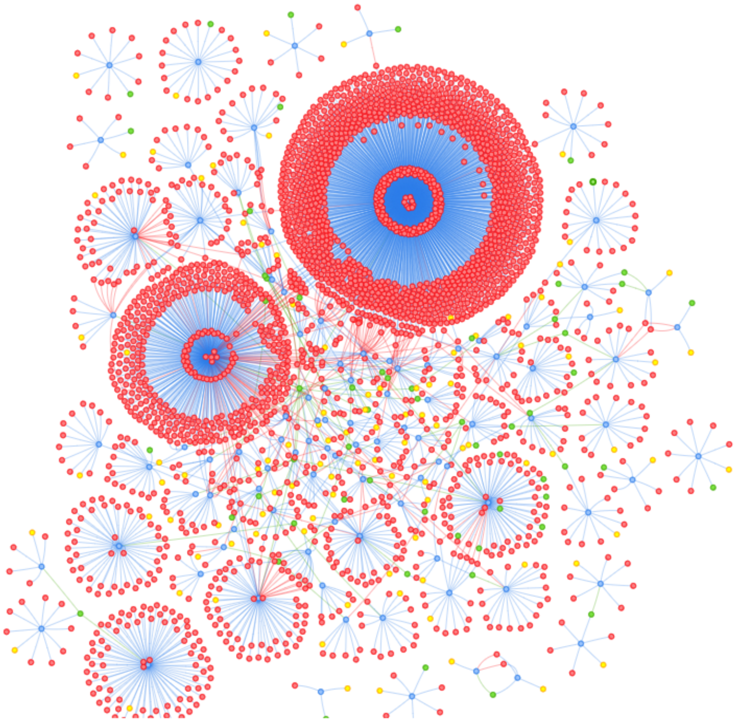

Fig. 11. Network of Likes range (501-1000) models with arXiv tags

5 Conclusion

This work present an insightful exploration of Hugging Face repository by structuring diverse features of machine learning models. As these machine learning models are crucial for the research communities and thus nourish the interdisciplinary research. However, modelling of heterogeneous features into information network will provide the potential linking among features and also pave to way for explore inter-repository scholarly information. The exploration also indicates the coverage of various computing libraries in interdisciplinary research activities. Furthermore the linkage to scholarly resources from arXiv is the key aspects and could be beneficial to compile available information against the models. As these cross repositories linkages are essential in exploiting the potential links of models to their corresponding research papers or potential references. In this

AI driven era, this could also provide the potential reference to context to researchers for a deep insightful.

This work could also be enhanced in future to incorporate the other scholarly repositories and also encapsulation of traditional metadata could be beneficial for communities and contribute towards effective implications of research data management and open science. Moreover, a potential direction could also be to incorporate the broader scholarly metadata features for a comprehensive network representation of metadata and their corresponding AI models which could be lead to recommender systems and relevant research community detection. The addition of spatial features will also provide the location based concept network which could provide region based publications focusing specific machine learning models.

Acknowledgments. This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), NFDI4DS (Grant number 460234259) and SmartER (Grant number 515537520). Authors also acknowledge the Hugging Face as data sources and also thanks the individuals involved in this research.

8

References

1. Ait, A., Izquierdo, J.L.C., Cabot, J.: Hfcommunity: An extraction process and relational database to analyze hugging face hub data. Science of Computer Programming **234**, 103079 (2024)
2. Castaño, J., Martínez-Fernández, S., Franch, X.: Lessons learned from mining the hugging face repository. arXiv preprint arXiv:2402.07323 (2024)
3. Castaño, J., Martínez-Fernández, S., Franch, X., Bogner, J.: Analyzing the evolution and maintenance of ml models on hugging face. arXiv preprint arXiv:2311.13380 (2023)
4. Castaño, J., Martínez-Fernández, S., Franch, X., Bogner, J.: 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). pp. 1–12. IEEE (2023)
5. Du, N., Guo, J., Wu, C.Q., Hou, A., Zhao, Z., Gan, D.: Recommendation of academic papers based on heterogeneous information networks. In: 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA). pp. 1–6. IEEE (2020)
6. Jiang, W., Synovic, N., Hyatt, M., Schorlemmer, T.R., Sethi, R., Lu, Y.H., Thiruvathukal, G.K., Davis, J.C.: An empirical study of pre-trained model reuse in the hugging face deep learning model registry. In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). pp. 2463–2475. IEEE (2023)
7. Liu, J., Shi, C., Yang, C., Lu, Z., Philip, S.Y.: A survey on heterogeneous information network based recommender systems: Concepts, methods, applications and resources. AI Open **3**, 40–57 (2022)
8. Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., Xia, F.: A survey of scholarly data visualization. Ieee Access **6**, 19205–19221 (2018)

9. McMillan-Major, A., Osei, S., Rodriguez, J.D., Ammanamanchi, P.S., Gehrmann, S., Jernite, Y.: 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 (2021)
10. McQuilton, P., Batista, D., Beyan, O., Granell, R., Coles, S., Izzo, M., Lister, A.L., Pergl, R., Rocca-Serra, P., Schaap, B., et al.: Helping the consumers and producers of standards, repositories and policies to enable fair data. Data Intelligence **2**(1-2), 151–157 (2020)
11. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
12. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Heterogeneous information networks: the past, the present, and the future. Proceedings of the VLDB Endowment **15**(12) (2022)
13. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
14. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: 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 (2020)