2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)

Answering User Questions about Machine Learning Models through Standardized Model Cards

Tajkia Rahman Toma, Balreet Grewal and Cor-Paul Bezemer
Electrical and Computer Engineering
University of Alberta, Edmonton, Canada
Email: {tajkiatoma, balreet, bezemer}@ualberta.ca

***Abstract*—Reusing pre-trained machine learning models is becoming very popular due to model hubs such as Hugging Face (HF). However, similar to when reusing software, many issues may arise when reusing an ML model. In many cases, users resort to asking questions on discussion forums such as the HF community forum. In this paper, we study how we can reduce the community's workload in answering these questions and increase the likelihood that questions receive a quick answer. We analyze 11,278 discussions from the HF model community that contain user questions about ML models. We focus on the effort spent handling questions, the high-level topics of discussions, and the potential for standardizing responses in model cards based on a model card template. Our findings indicate that there is not much effort involved in responding to user questions, however, 40.1% of the questions remain open without any response. A topic analysis shows that discussions are more centered around technical details on model development and troubleshooting, indicating that more input from model providers is required. We show that 42.5% of the questions could have been answered if the model provider followed a standard model card template for the model card. Based on our analysis, we recommend that model providers add more development-related details on the model's architecture, algorithm, data preprocessing and training code in existing documentation (sub)sections and add new (sub)sections to the template to address common questions about model usage and hardware requirements.**

***Index Terms*—Machine learning model hubs, Model cards, Questions & answers, Hugging Face**

I. Introduction

Reusing pre-trained machine learning models has become very popular in recent years as it reduces the cost and development complexity of training a model from scratch [19], [51], [23]. One increasingly popular way to access these pre-trained models is through model hubs, such as Hugging Face (HF) [22], [26], [27]. When users encounter problems with, or have questions about the models, they can turn to the community on these hubs for help. For example, HF provides a platform for users to communicate with the model community through HF discussions [40] inside model repositories, which are similar to issues on GitHub. Unfortunately, answering these questions is done by volunteers from the model community, and often, questions are left unanswered, which in turn blocks the questioner from using the model. In this paper, we study what types of questions are being asked, and we investigate whether they could have been answered through improved model documentation, which would both reduce the question-answering burden and the waiting time for an answer.

Many studies have worked on helping answer questions in question-answering communities like Stack Overflow. They provide suggestions for information seekers and platform designers on how to get a faster answer [57], [55]. Some studies improve platform design, such as automatic tag recommendation [56], highlighting content [35], [3] or ranking candidate answers [1] to help users find relevant results. To the best of our knowledge, our study is among the first to use model documentation to answer user questions on machine learning (ML) models to improve the user experience on ML model hubs. The purpose of model documentation is to help model users understand and effectively utilize ML models [27]. We study empirically how to optimize the use of model documentation based on the questions being asked about the models.

In this study, we analyze 11,278 user questions from HF discussions to understand questions related to ML models and analyze how we can effectively address these questions. We examine discussion metadata to measure community effort in addressing the questions. Through topic modeling, we identified the high-level topics of discussions that contain questions to understand their underlying nature and investigate the possibility of addressing the questions in model documentation. Finally, through manual analysis, we explore how standard model documentation (such as a model card template [5], [15], [34]) can answer questions faster than waiting for a response. The study addresses the following research questions:

- **RQ1: How much effort is spent on handling questions?** We analyzed the HF community's effort in responding to questions to understand the time involved. Our analysis indicates that it does not cost the community much effort to respond to questions. However, 71.9% of the questions are still open, of which, 40.1% are open without any response. Using model documentation to answer questions can help the community focus more on the remaining open questions.
- **RQ2: What are the high-level topics of questions containing discussions?** To understand the types of questions asked, we clustered high-level topics of posts. We identified 5 clusters encompassing 31 topics related to various aspects of model development, model usage, and requests for model variants. The findings suggest that most questions and their underlying topics can be

1558-1225/25/$31.00 ©2025 IEEE
DOI 10.1109/ICSE55347.2025.00066

answered through the model documentation, as they touch upon common topics.

- **RQ3: Which questions could be answered using a standard model card template?** We align questions asked with (sub)sections of model cards to assess how effectively a standard model card template can provide information. We found 42.5% of the analyzed questions could be answered through following a standard model card template, showing the template's usefulness in addressing questions. For questions that cannot be answered with a standard model card template (unmapped questions), we found that adding (sub)sections on how to use the model and its hardware requirements can address some of the common unmapped questions.

Our study reveals that following a standard model card template can help address many questions. Hence, we encourage model providers to adopt a standard template for their model cards. Furthermore, our study sheds light on a discrepancy between the questions asked and the information provided in current model cards indicating a need for further enhancements to the template. We suggest model providers to include more development-related details in the model cards, and add additional (sub)sections on model usage and hardware requirements to answer common user questions.

The remainder of the paper is structured as follows. Section II provides background knowledge to increase the readability of this paper. Section III discusses our study setup. Sections IV, V and VI discuss the results of our study. Section VII discusses the implications of our study. Section VIII discusses the threats to the validity of this work and Section IX summarizes the related work. Section X concludes our work.

II. Background

A. HF Model Discussions

HF provides a central hub for developers to publish their models along with a wide range of datasets for model training and evaluation [40]. It also offers tools for building and deploying applications on top of these models. A key new feature (added in 2022) of HF is the community tab in each repository, allowing community contributions through pull requests and discussions[1]. The discussions enable community members to ask and answer questions about a particular model, and share ideas and suggestions directly with the repository owner and the community[2].

A *repository owner* can be an individual or an organization that owns a repository to a particular model. An organization may have one or more team members working on its repositories, referred to as *organization team members*. Together, we refer to the *repository owner* and *organization team members* as *model providers*. Further, the community of *model providers* and model users are referred to as the *model community*. We refer to the discussions made about a model simply as *discussions* throughout the paper. HF also provides a community discussion forum[3] where users, developers and researchers can engage in discussions on different categories of topics that are not necessarily model-specific. Our study is limited to the analysis of discussions which comes from only repositories of models and not the forum.

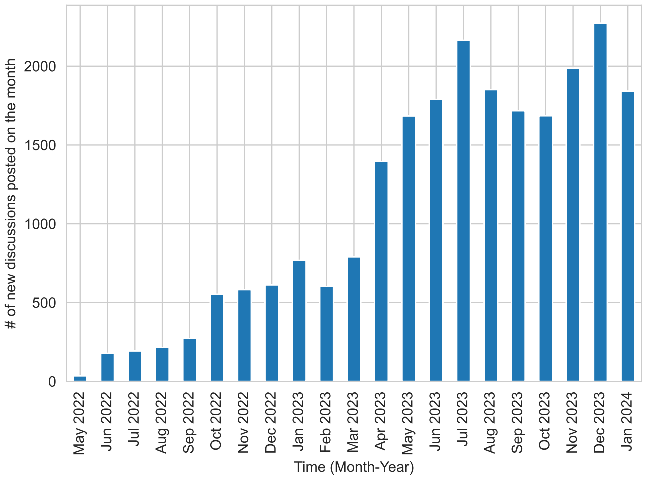


Fig. 1: The number of new discussions over time

Like the increase in the number of models, with a weekly growth rate of 3.16% [27], the number of discussions is also increasing. Figure 1 shows the increase of the number of new discussions in each month from May 2022 to January 2024 with an average monthly growth of 1,104 discussions.

B. HF Discussion Components

A discussion in HF is a simpler version of those found in other git hosts, like Issues on GitHub[1]. Here, we explain and name each part of a HF discussion. Figure 2 is an example of a HF discussion[4]. It consists of the following:

1) **Title**: A brief overview of the topic being discussed.
2) **Discussion Status**: Indicates whether the discussion is open or closed. As HF discussions are based on GitHub Issues, we can derive that every discussion is open unless it is either answered or not relevant/cannot be answered[5].
3) **Post**: A detailed explanation of the question or issue being addressed or discussed in the discussion.
4) **Post Author**: The author of the post, who initiates the discussion or asks the question.
5) **Post Time**: The date and time when the discussion was posted or last updated.
6) **Response**: The suggestion or thought shared by other community members about the discussion.
7) **Response Author**: The author of the response, who contributes to the discussion with an answer or insight.

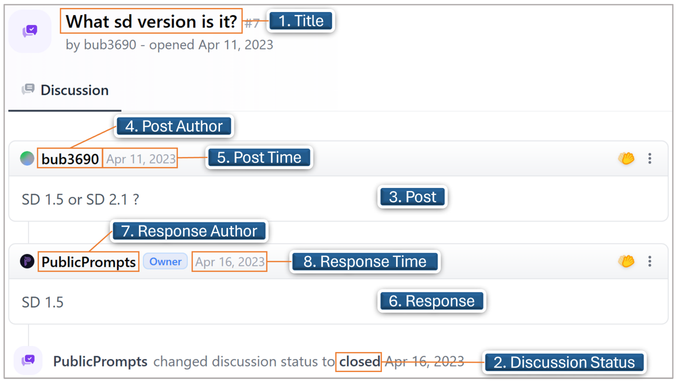
[1] https://huggingface.co/docs/hub/repositories-pull-requests-discussions

[2] https://huggingface.co/blog/community-update

[3] https://discuss.huggingface.co/

[4] https://huggingface.co/PublicPrompts/All-In-One-Pixel-Model/discussions/7

[5] https://docs.github.com/en/issues/tracking-your-work-with-issues/closing-an-issue



What sd version is it? #7

by bub3690 - opened Apr 11, 2023

Discussion

bub3690 Apr 11, 2023

SD 1.5 or SD 2.1 ?

PublicPrompts Owner Apr 16, 2023

SD 1.5

PublicPrompts changed discussion status to closed Apr 16, 2023

Fig. 2: Overview of the components of an HF discussion[4]

8) **Response Time**: The date and time when the response was provided to the post.

III. Study Setup

Our goal is to analyze questions from the HF community about models to help answer questions faster. First, we collected the list of models and their discussions from the HF community and filtered the data to ensure quality. Then we identify the discussions that contain a question using `GPT-3.5 Turbo` for our study. Figure 3 outlines the steps of our data collection and preparation. We collected our data in January 2024. In the remainder of this section, we explain each step in detail.

A. Step 1: Collecting Data

1) Collect List of Models: We first collect the list of models from HF so that we can access the corresponding discussions. We collect the list of models utilizing the HF Hub API[6], which produced a total of 480,691 models.

2) Collect Model Community Discussions: We collected the discussions for each model using the HF Hub's Community API[7]. These discussions also include pull requests. Since we focused solely on analyzing the questions in the discussions, we discarded the pull requests (93,463 in total) from our collected data. Out of the 480,691 models, we found 23,188 discussions from 9,211 models, with the remaining models not having any discussions. The discussions span from May 2022 to January 2024.

B. Step 2: Filtering Data

1) Filter Models: To reduce the number of spam or toy models in our analysis, we filtered the models by their number of downloads and likes. We kept the models that have at least 1 like and 1 download. The filtering process resulted in a final list of 29,781 models.

2) Filter Discussions: To identify the criteria for filtering out discussions, we performed a preliminary analysis of a sample discussion. We studied a randomly selected statistically representative sample of discussions with a 95% confidence level and 5% error margin from 23,188 discussions (378 discussions). Based on our analysis, we listed the following criteria to filter out the data.

- **Hidden discussions:** Some discussions had posts made private, resulting in the posts not being publicly visible. Consequently, we removed discussions with hidden/private posts.
- **Short discussions:** Some discussions were too short or incomplete to comprehend. After examining them manually, we discovered that discussions with less than 50 characters lack meaningful content. Therefore, we excluded such discussions.
- **Non-English discussions:** Certain discussions were not in English, making it difficult for us to grasp the content. Consequently, we excluded these non-English discussions from our analysis utilizing the xlm-roberta-base-language-detection [29] model.

We applied the filters on all discussions to get a refined list for further analysis, resulting in 15,964 discussions.

C. Step 3: Identifying Questions

To identify the discussions that contain a question, we utilized `GPT-3.5 Turbo` due to its renowned natural language processing capabilities [13], [21]. To determine how GPT performs in our case, we first performed a sample analysis.

1) Classify Sample Discussions using GPT: We applied the filters from Section III-B2 and classified the remaining 331 sample discussions using the prompt specified in Figure 4. We added the 2^{nd} and 4^{th} paragraph in the system prompt to decrease the number of false-negatives (GPT chooses "no", our agreed class is "yes") and false-positives (GPT chooses "yes", our agreed class is "no") respectively. For each discussion, we replaced the `<title of the discussion>` from the user prompt with the title of the discussion and `<content of the post>` with the content of the post. We executed the `gpt-3.5-turbo-0125` classifier three times for each discussion with the prompt with a temperature of 1, and considered the majority class ("yes" or "no") as the final class.

2) Evaluate Performance of GPT as Classifier: We evaluated GPT's ability to identify if a post contains questions. Two authors manually reviewed the 331 discussions individually and labeled them as containing a question or not to prepare the ground truth. They achieved a high Cohen's Kappa coefficient of 0.94. The labelling disagreements were resolved through discussion and consensus. For 4 discussions, where disagreements could not be resolved, the third author acted as a tiebreaker. `GPT-3.5 Turbo` achieved a high accuracy of 94% and a high F1 score of 95%, demonstrating its robustness and reliability in distinguishing posts with questions.

We used `GPT-3.5 Turbo` to classify all 15,964 discussions that passed the cleaning filters. We encountered errors classifying 15 discussions due to their long token length or to

[6] https://huggingface.co/docs/huggingface_hub/en/package_reference/hf_api

[7] https://huggingface.co/docs/huggingface_hub/en/package_reference/community

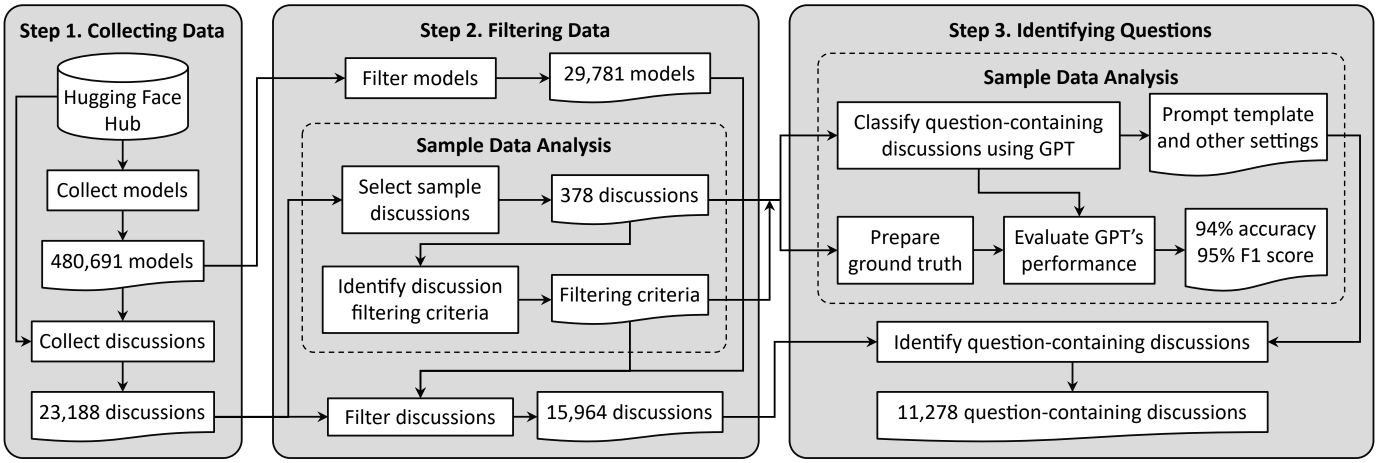


Fig. 3: Overview of our process to collect discussions containing questions

```
System: You will be presented with a discussion of a Hugging
Face model. You will answer the following question with ‘‘yes” or
‘‘no” for the discussion:
Have any questions been asked in the discussion?

Think sentence by sentence. Be aware that there can be no ‘‘?”
mark at the end of the question.

Please give your answer in the following format:
contains_question: yes/no

If the answer is ‘‘yes”, return me the part of the discussion that
contains the question in the following format:
question_part: part of the discussion containing the question
―――――――――――――――――――――――――――――――――――――――――――――――――――
User: <title of the discussion>

<content of the post>
```

Fig. 4: Prompt template used to classify each discussion

having many repetitive letters or symbols like ' or \n in the discussions, which we excluded from our analysis. We ended up with a total of 11,278 discussions that contained questions. Our collected data and analysis are available in our replication package [52].

IV. RQ1: How much effort is spent on handling questions?

Motivation: The increasing use of HF models has led to an increasing number of model community discussions. This research question aims to assess the effort invested in responding to questions to better understand the time commitment invested by HF model community members.

Approach: We measure the effort spent on handling questions in terms of (1) the number of responses to each question, (2) the length of the responses, (3) the number of unique participants in each discussion and (4) the time between posting the question and receiving the first response.

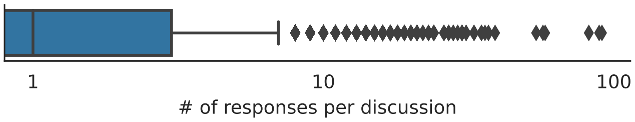


Fig. 5: Distribution of number of responses per discussion

We used the Mann-Whitney U Test [33] to compare the distributions of the number of responses per discussion based on (1) discussion status and (2) responses from model providers. The Mann-Whitney U Test is a non-parametric statistical test used to check if two non-normal distributions differ. A p-value below 0.05 indicates a significant difference between the distributions. We calculated the effect size of the statistical difference between the distributions using Cliff's Delta to measure the magnitude of the difference. We used the thresholds proposed by Romano et al. [44] to interpret the delta value d: negligible if $|d| \leq 0.147$; small if $0.147 < |d| \leq 0.33$; medium if $0.33 < |d| \leq 0.474$; and large if $0.474 < |d| \leq 1$.

Findings: **Among the 11,278 discussions containing questions, 7,782 discussions have at least one response.** A discussion with a question gets a median of one response. Figure 5 shows that the number of responses per discussion ranges from 0 to 91. With a Mann-Whitney U Test, we found that the distribution of the number of responses varies noticeably (with a small effect size of -0.33) based on the discussion status, meaning that open discussions have fewer responses than closed discussions. From Figure 6, we see that open discussions receive a median of one response, whereas, closed discussions receive a median of two responses. We notice that open discussions are left unanswered or contain fewer responses compared to closed discussions.

Responses to discussions containing questions have a median length of 24 words. Figure 7 shows that the length of responses varies from 0 to 20,216 words (which were mostly made up by the contents of error log that was included in the response). From Figure 8, we see that responses for closed

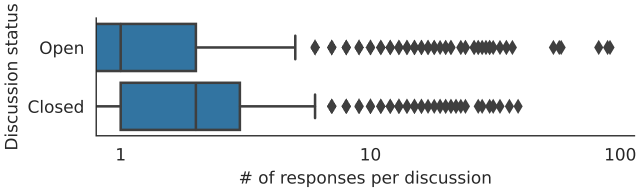


Fig. 6: Comparison of number of responses per discussion based on discussion status

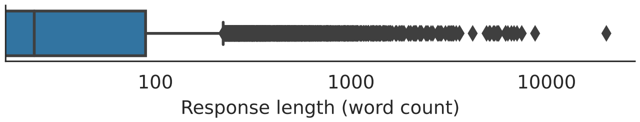


Fig. 7: Distribution of the number of words per response

discussions are longer than open discussions. Open discussions receive responses that are a median of 14 words long, while closed discussions receive responses that are a median of 45 words long.

A discussion with a question has a median of two participants, the questioner and another person. The number of participants in discussions ranges from 1 to 54, as shown in Figure 9. A Mann-Whitney U Test shows that the number of responses varies significantly with a large effect size of 0.63 based on the participation of model providers in discussions.Figure 10 shows that discussions with at least one response from a model provider receive a median of two responses, while those without typically get none. Additionally, Figure 11 shows that discussions are much more likely to be closed with a model provider's participation. The number of responses and status of the discussions with a model provider's participation suggest that the questions require specialized knowledge about the model to answer.

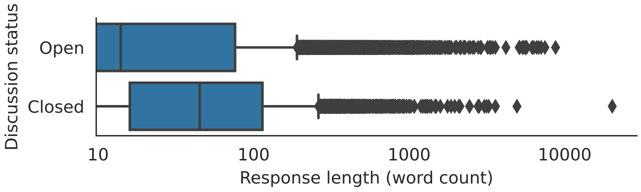


Fig. 8: Comparison of the length of the responses based on discussion status

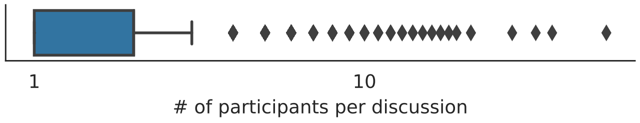


Fig. 9: Distribution of the number of participants per discussion

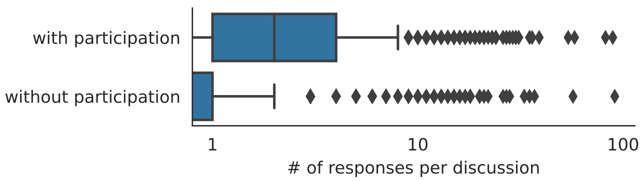


Fig. 10: Comparison of the number of responses per discussion based on the participation of a model provider

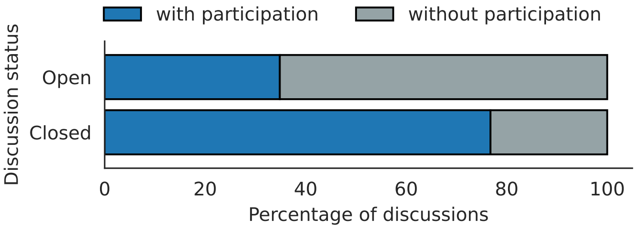


Fig. 11: Comparison of discussion status based on the participation of model providers in a discussion

A median discussion containing a question gets a response on the same day. The response delay ranges from 0 to 453 days. Among the 7,782 discussions that got a response, 4,820 were responded to within 24 hours, making up 42.7% of the total question-containing discussions. However, **many open discussions did not receive a response at all**. Among the 3,496 discussions that did not get a response, 3,252 are open discussions, making up 28.8% of the total question-containing discussions, of which 3,236 have been open for more than 24 hours.

> Questions are typically handled by model providers and responses are typically short. If a question does not get a response within 24 hours, it is unlikely to ever get a response. 71.9% of the question-containing discussions are open, of which, 40.1% are open without any response.

V. RQ2: What are the high-level topics of questions containing discussions?

Motivation: To understand the types of questions users ask about models, this research question identifies the high-level topics of the posts. We use automated topic analysis to investigate whether it is possible to address the questions in the model documentation from a high level. Knowing prevalent

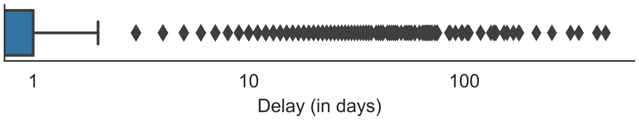


Fig. 12: Distribution of response delay per discussion

discussion topics enables model providers to organize the documentation accordingly.

Approach: The goal of this research question is to determine whether the questions relate to topics that could be included in the model card. Therefore, we categorize all the questions based on their topic. We use BERTopic [31], a topic modelling technique, to identify topics in the studied discussions. BERTopic has recently been gaining attention to identify discussion topics [2], [17], [20], [16], [47], [54], [12]. It provides more meaningful and diverse topics than other popular techniques like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) or Top2Vec [17], [18], [48], [42].

We prepared the discussion posts that contained questions for topic analysis by removing unnecessary elements such as code blocks, images, emojis, and URLs, which do not help BERTopic understand the content. Since the BERTopic documentation does not recommend any data preprocessing, we did not further process the discussions. BERTopic consists of five steps: embedding, dimensionality reduction, clustering, tokenization, and weighting. Each step is modular and can use different algorithms. Following BERTopic's best practices[8], we used SentenceTransformer for embedding, UMAP for dimensionality reduction, HDBSCAN for clustering, CountVectorizer for tokenization, and c-TF-IDF for topic representation. We also used `GPT-3.5 Turbo` to assist in labelling the topics with short, representative names based on a prompt template from the BERTopic documentation. One author further went through the representative documents of the topics (which are automatically selected by BERTopic) manually to fine-tune the labels. The author used the GPT-generated labels as a guide and adjusted the labels based on their own observations of the topic keywords and representative posts.

By setting BERTopic to cluster a minimum of 60 posts per topic, we identified 31 meaningful topics and one outlier topic. This threshold of 60 posts per topic was defined empirically through trial and error to produce a meaningful output. We further grouped similar topics to simplify the analysis and clarify the main areas of interest. We used Hierarchical Topic Modeling[9] with BERTopic allowing topics with a distance of less than 1 to merge, condensing the 31 initial topics into 5 clusters. The distance among topics and a visualization of how they were grouped into clusters are available in our replication package [52]. We labelled the clusters similar to how we labelled the topics.

Findings: **We identified 5 clusters containing 31 topics from posts with questions**. Table I lists these clusters with their topics along with the number of posts in each topic, a short label for each topic and their top 10 keywords. Here, we briefly describe each cluster based on its topics' representative documents and keywords.

Cluster 1: Challenges with Model Setup and Optimization (2,019 questions) The issues discussed in this cluster cover various aspects of setting up and optimizing models. Users are mainly concerned with the context length and maximum input capacity of the model, as well as the tokenizer files like tokenizer.json or tokenizer_config.json. Another major focus is questions about different natural language support by models. The cluster also contains questions about prompt templates and composition of prompts. Finally, the discussions highlight performance challenges, such as slow inference times, and issues with transcribing and processing audio files. We believe the common questions on topics like maximum context length, tokenizer files, multilingual support, prompt template formatting and performance challenges can be addressed in the model documentation to provide users with a comprehensive guide. Some discussions also focus on the analysis of the model's behaviour with Not Safe For Work (NSFW) contents.

Cluster 2: Errors and Issues with Model Loading, Usage and Deployment (1,687 questions) The posts in this cluster focus on issues related to loading, using, and deploying models across different environments like HF Transformers, local machines and cloud platforms (e.g. AWS SageMaker). Common problems include setup errors, missing dependencies, memory and VRAM limitations, GPU compatibility issues, and API errors during inference. There are also issues with loading models through web interfaces (e.g. oobabooga). The cluster shows the technical challenges and troubleshooting needed for model use in various environments and suggests that comprehensive documentation could help resolve some of the issues.

Cluster 3: Seeking Assistance with Model Usage (1,503 questions) This cluster covers details on requesting and providing guidance on the usage of the models. Some posts provide guides on training and optimizing Stable Diffusion, an AI model for generating images. There are also requests for assistance on topics like license details for commercial use, model download errors, and using and citing the model. Additionally, users request help converting and exporting the model to ONNX format. The requests for help on model usage scenarios indicate a need for more detailed documentation and support resources for users to effectively utilize the models.

Cluster 4: Questions about Model Training, Evaluation and Fine-Tuning (1,226 questions) The posts in this cluster discuss various challenges and techniques related to the training and fine-tuning of the models. Several posts inquire about the datasets used for training and evaluating models, including requests for dataset descriptions and evaluation scores. Other documents focus on fine-tuning the models to suit unique use cases or incorporate new datasets. Furthermore, there are queries concerning the sharing and interpretation of training codes and results, particularly the hyperparameters and evaluation methods employed. Finally, there are technical questions regarding issues encountered during fine-tuning. The questions in this cluster highlight user interest in more details on the dataset and code used for model training and evaluation, as well as fine-tuning the model for their own use cases.

[8] https://maartengr.github.io/BERTopic/getting_started/best_practices/best_practices.html

[9] https://maartengr.github.io/BERTopic/getting_started/hierarchicaltopics/hierarchicaltopics.html

Cluster 5: Updates and Requests for Model Versions (1,053 questions) This cluster predominantly discusses users' questions about future plans, specific model merges, and comparisons among similar models. Users are requesting various versions of the models, particularly in the GGUF format, a binary file format for quantized models, highlighting the need for specific versions to overcome hardware limitations. Users also ask for a model variant with a different number of model parameters, indicating the need for more parameters for a better-performing model and fewer parameters for lower computation resources. There are also questions about the differences between the two models, indicating a clear need to better distinguish between models. We also find discussions on Mistral-7B, a large language model created by the AI research company Mistral, indicating a high user interest in the model.

> The question-containing discussions are centred on different topics of model development, model usage, and requests for model variants. Regarding model development, we find users asking for details about model training and evaluation, and assistance for fine-tuning. Regarding model usage, we find users facing challenges in different phases of model loading, setup, inference, deployment and usage in general. Our findings highlight that many of these topics can be anticipated and should be addressed better in the model documentation.

VI. RQ3: Which questions could be answered using a standard model card template?

Motivation: With the increasing number of questions posted in HF discussions, one way to reduce the time model providers spend answering questions and to give answers faster is to integrate common questions' answers into the model documentation. In this research question, we identify whether the current industry standard template of a model card can answer the questions asked. By mapping questions from discussions to the standard model card template, we aim to identify (1) how useful it would be for model providers to follow the template and (2) improvements to the template.

Approach: A model card is part of a model's documentation that serves the purpose of providing detailed information about the models to facilitate user understanding and increase the transparency of the model to its stakeholders [5], [15], [34]. Many recent studies [8], [37], [27], [36] use the model card template of Mitchell et al. [34] as a standard. Therefore, we also adopt this template. The model card template includes 9 sections: Model Details, Intended Use, Factors, Metrics, Evaluation Data, Quantitative Results, Ethical Considerations, Broader Impacts, and Caveats. Some sections contain subsections to describe specific details of the section, having a total of 27 subsections for a model card.

We mapped user questions to the most relevant (sub)sections of the template based on what was asked. To be more precise to map the questions to model card (sub)sections, we manually analyzed the questions. With a 95% confidence level and a 5% margin of error, we randomly selected a statistically representative sample of 372 discussions from the total 11,278 question-containing discussions. We manually recorded the questions from these posts. Out of 372 discussions, 359 contained 419 questions, with 44 posts having multiple questions. Thirteen posts did not include any user questions, however, they were marked as such either because they discussed the model's question-answer capabilities or due to GPT's error in detecting questions from the post.

We then mapped each of the 419 questions to the appropriate (sub)section of the model card template, based on where the answer would be found in the standard model card. Initially, two authors (one PhD student with 3+ years of professional experience, one MSc student) completed a closed card sort on the first 160 questions based on their experience to establish a better mutual understanding of the model card template. Then the two authors individually mapped the next 100 questions, achieving substantial agreement with Cohen's Kappa coefficient of 0.73. After comparing and resolving disagreements through discussion, the authors mapped the remaining 159 questions, reaching an inter-rater agreement of 0.81, indicating a very good agreement.

Findings: **We could map 42.5% (178) of the total 419 questions to a (sub)section of the model card template.** The questions were mapped to 16 subsections found in 6 of the 9 sections of the model card template. Table II shows the distribution of questions across the different sections and subsections.

Most of the questions (49.4%) were related to the `Model Details` section, showing a strong interest in a model's technical details. Users asked 28 questions related to the `Training detail` of the model to understand the model's training process, including the code used, training parameters, and the quantization process for quantized models. We also noticed many questions (22) relating to the `Paper or other resource` of the model. Repeated questions were asked about the code or script used to train or fine-tune the model or how to use the model. Thus, providing training or the model's usage instruction code in the model card can improve the model's reproducibility and usability.

Furthermore, we identified repeated questions on `License` in three groups: whether the model can be used for commercial purposes, details about the current license, and if the license allows redistribution of the model in different platforms. This shows the importance of providing clear and detailed licensing information in the model card. Easily accessible license details can help users decide if the model is suitable to use in specific cases. Additionally, users repeatedly asked about the context size or number of tokens to use in the model, which we mapped to the `Model type` subsection. The `Model type` section also includes questions about the model's architecture and specifications, showing users' interest in its technical details and constraints.

We then have 21.4% of questions, the second highest, from the `Intended Use` section of the template. While the `Primary intended uses` subsection typically outlines scenarios for using the model, we found that user ques-

TABLE I: List of the 31 topics in 5 clusters generated using BERTopic. The topics of the same cluster are grouped together

Topic Label	Top 10 Keywords	# of Posts
Cluster 1: Challenges with Model Setup and Optimization		**2,019**
Maximum Context Length	length, context, tokens, context length, max, maximum, 2048, input, model, text	515
Tokenizer File	tokenizer, token, json, tokens, tokenizer model, model, tokenizer_config, tokenizer_config json, special, error	385
Multilingual Support	language, english, languages, chinese, translation, model, french, support, multilingual, translate	349
Prompt Design	prompt, prompt format, format, template, prompt template, assistant, user, instruction, prompts, use	276
Multilingual Audio Transcription	audio, whisper, transcription, speech, transcribe, voice, language, file, output, model	224
Optimizing Inference Time	electronic, music, inference, speed, time, td, slow, faster, inference time, inference speed	141
Prompt for Question Answering	question, answer, answering, question answering, context, answering model, prompt, model, answers, document	66
NSFW Censorship Analysis	nsfw, uncensored, censored, scene, censorship, life, meaning, rules, oh, say	63
Cluster 2: Errors and Issues with Model Loading, Usage and Deployment		**1,687**
Model Loading Error	layers, model layers, model, transformers, self_attn, error, bias, py, file, from_pretrained	523
GPU Memory Allocation Issue	memory, gpu, model layers, layers, self_attn, g_idx, g_idx model, bias model, model, vram	521
Inference API Error	inference, api, inference api, 2023 07, 07, 2023, huggingface, 43, 42, endpoint	258
Model Loading Error in WebUI	webui, generation webui, text generation, generation, file, py, py line, text, line, oobabooga_windows	241
Model Deployment Error	sagemaker, deploy, endpoint, aws, error, tgi, inference, aws sagemaker, instance, ml	144
Cluster 3: Seeking Assistance with Model Usage		**1,503**
Stable Diffusion Web UI Tutorial	diffusion, stable, stable diffusion, image, ui, web ui, web, images, automatic1111, size	843
Model License	license, commercial, apache, mit, commercial use, use, license model, commercially, model, license hi	310
Model Download Issues	download, download model, model, git, files, error, locally, run, load, access	135
How to Use Model	use, use model, example, model, app, usage, help, code, guide, example code	80
DOI Request	doi, cite, request doi, paper, request, citation, model paper, doi hi, work, like cite	72
ONNX Model Export	onnx, onnx model, model onnx, export, convert, onnx version, onnxruntime, model, js, optimum	63
Cluster 4: Questions about Model Training, Evaluation and Fine-Tuning		**1,226**
Model Training Dataset	dataset, data, bits, used, training, dataset used, int, set, std, binary	353
Sentence Embeddings for Semantic Similarity	embeddings, similarity, sentence, score, embedding, model, label, sentences, use, word	207
Model Evaluation and Benchmarking	evaluation, score, leaderboard, scores, benchmarks, results, accuracy, model, benchmark, humaneval	186
Fine-tune Model on Dataset	fine, tuning, fine tuning, tune, fine tune, tune model, model, model fine, tuning model, dataset	155
Training Code Sharing	training, code, training code, share, share training, script, training script, train, scripts, thanks	136
Geneformer Perturbation Error	gene, cell, geneformer, perturbation, data, cells, dataset, insilicoperturber, silico, output_directory	102
LoRa Model Fusion	lora, loras, fine, lora model, lcm, training, peft, model, lora fine, tuning	87
Cluster 5: Updates and Requests for Model Versions		**1,053**
GGUF Model Version Request	gguf, version, ggml, gptq, quantized, quantization, model, bit, llama, gguf version	616
Different Parameter Version Request	13b, version, 7b, model, 3b, 70b, 33b, plans, 13b version, 30b	166
Model Merging Process	merge, merging, merged, models, mergekit, did, model, merge models, did merge, adapter	102
Model Discrepancy Investigation	llama, llama2, chat, teacher, llama 7b, older, 7b, old, years, mother	98
Mistral-7B Fine-Tuning	mistral, mistral 7b, 7b, solar, v0, 7b v0, mistral model, 10 7b, solar 10, mixtral	71
Outlier	-	3,790
Total		**11,278**

tions focused more on technical details, such as whether the model output can be adapted for their specific usage scenarios. Therefore, alongside describing usage scenarios, the model card template should include information on how flexible the model output is for different applications.Additionally, we encountered 7 questions on `Out-of-scope use cases` where users wanted to understand the differences between the model and others. We also noted 3 posts where users sought advice on selecting models based on their requirements.

The third highest portion (15.2%) of questions are about the `Training Data` section. Users were keen to know about the source and composition of the training data, which should be addressed in the `Datasets` subsection. Additionally, questions were raised about how the training data was prepared, highlighting the need for detailed information on the dataset and preprocessing methods. This transparency is crucial for assessing the model's generalizability and reliability based on its training data.

To determine whether users asked questions despite the answers being in the model card, we examined the model card version from before the questions were asked. **We found 22 questions were asked, even though the answers were in the model card.** Some questions can be answered through the model tags, and some are on the base model's model card. This shows the need to make the model cards more navigable and searchable so that users can quickly find the information they need.

From the unmapped questions, we found repeated ques-

TABLE II: Our mapping of user questions to model card template sections and subsections

Sections	Subsections	# of Questions
Model Details	Training detail	28
	Paper or other resource	22
	License	21
	Model type	16
	Citation details	1
		(49.4%) 88
Intended Use	Primary intended uses	28
	Out-of-scope use cases	10
		(21.4%) 38
Training Data	Datasets	20
	Preprocessing	5
	Motivation	2
		(15.2%) 27
Quantitative Analyses	Unitary results	12
	Intersectional results	1
		(7.3%) 13
Evaluation Data	Preprocessing	6
	Datasets	2
		(4.5%) 8
Metrics	Variation approaches	3
	Model performance measures	1
		(2.2%) 4
Mapped Total		**(100%) 178** **42.5% of total 419 questions**

tions on topics that can be added as (sub)sections in the model card. We noted the context of the 240 unmapped questions, focusing on the subject of the question, and grouped them accordingly. **We found a significant number of questions (82) about how to use the model** covering topics from model input to model loading, model usage and model output. Questions about model input are mainly focused on prompt or instruction formats. Questions about model loading are mostly requesting instructions on how to load the model. Questions about model usage generally inquire about using the model overall, with a UI or through the HF Inference API, across different operating systems, and on a GPU. This emphasizes the challenges users face at various stages of model use. A (sub)section on how to use the model could address these concerns and improve user experience.

Furthermore, we found 19 questions about the memory or hardware requirements to load, deploy, or fine-tune the model, showing user concerns about the computational resources needed for effective use, deployment or fine-tuning. Offering detailed information on system requirements and hardware configuration recommendations in the model card can help users prepare better for deploying or fine-tuning the model in their specific computing environment.

We also found many questions (36) about potential future updates or enhancements to the model. Most inquiries are about a quantized version of the model (9) or the same model with different parameters (7). This reflects user interest in the ongoing development and improvement of the model. Addressing these questions in a roadmap or future development (sub)section could manage user expectations.

Our analysis also showed that some questions cannot be answered solely through the model card. For instance, 41 questions were about specific errors or bugs encountered while using the model, which are often user-specific cases. This supports our proposal of a comprehensive tutorial or documentation on effectively using the model, alongside the model card, to provide basic guidance to users. We also encountered 16 questions regarding model fine-tuning and 9 questions asking for training instructions to train a similar model with a different dataset. Since this process varies depending on scenarios and datasets, providing technical details on the model's training process and basic fine-tuning guidance would help users get started.

> Users ask more technical questions than what the model card currently covers. For unmapped questions, adding (sub)sections on using the model effectively and its hardware requirements for loading or fine-tuning the model can address recurring inquiries. We also noted questions about errors, bugs, and fine-tuning that may not be fully addressed by the model card alone. However, providing a detailed tutorial or user guide alongside the model card can address these issues and enhance the overall user experience with the model.

VII. Discussion

The number of discussions in HF is rapidly increasing. Although we do not find much effort involved in responding to questions in these discussions, our results indicate the feasibility of answering questions through model documentation. Proactively answering common and straightforward questions in model cards can reduce the number of questions posted and allow the community to focus on more complex, user-specific issues. From the topic analysis of the HF question-containing discussions, we find that they mainly focus on different topics of model development, model usage, and requests for model variants, many of which could be found in the model documentation. The analysis also suggests that the topics are primarily centred on the technical details of models, which can be best addressed by model providers.

From our manual analysis, we found that many questions can be answered following the standard model card template. However, the current model card template gives a general overview of the model. We suggest adding more details related to development in the subsections related to `Primary intended uses`, `Paper or other resource`, and `Training detail` of the model. We also propose adding (sub)sections to the template to address common questions about hardware requirements and model usage. In addition to our manual analysis, the thematic analysis of the discussions containing questions revealed 521 discussions about GPU memory allocation issues and 80 about how to use the model, supporting our proposal.

Bhat et al. [8], in their study about model documentation practice found that the `Model details` section is the most filled section in model cards, yet we still found many questions belonging to this section in our analysis, indicating support of their claim that the information in model documents is often vague. Their study also revealed that model documentation is often not self-contained and directs readers to additional resources. We also found user questions where the answers were in additional resources.

We also found questions from users after getting different errors. To address the errors and increase users' technical understanding of models, it may be beneficial to include supplementary documentation, such as tutorials or detailed technical guides, in addition to the model card. This will increase the transparency of model usage and help model users understand how to troubleshoot any issues, making the model more accessible and usable.

VIII. Threats to Validity

Internal Validity: The data filtering choices we made in Section III-B could affect the internal validity of our study. For instance, we only included repositories with at least one like and one download to eliminate spam or toy repositories. Our data shows that 91% of the models on HF have 0 likes, and 71% have 0 downloads. By using one like and one download as thresholds, we discarded 450,910 models with 5,332 discussions in total. A random check showed that most of these were toy or spam models. Selecting higher thresholds would discard many more discussions proportionally (e.g., 2 likes and 10 downloads would discard approximately 1,100 more discussions for 16,906 models). Future research should explore other data filtering methods on the HF model hub to ensure robust results.

Construct Validity: Effort is a difficult concept to measure. There is no single metric to capture effort; hence, we used four metrics as proxies to capture several aspects related to effort. However, these proxies may not fully capture the true meaning of effort. Future studies could use alternative metrics to better assess effort.

We relied on the status of discussions to determine if questions were resolved. Since there is no notion of 'answer' or 'accepted answer' for HF discussions, and the concepts of 'open' and 'closed' are not well defined, we derive the definitions from GitHub Issues and modified them to fit the context on HF. Future studies should consider alternative ways to determine if a question has been answered, like analyzing the discussion responses.

For topic modelling, BERTopic can yield different results depending on the data preprocessing step and training parameters. Furthermore, the cluster of topics can vary depending on BERTopic's clustering parameters. Therefore, we experimented with various discussion preprocessing steps, training and clustering parameters. We selected the parameters that produced the best results while minimizing interference. In addition, we labelled topics and clusters using our knowledge of model development disciplines and GPT assistance. Although the labels might not be perfect, we supported our choices with a manual review of representative posts to ensure they roughly describe the topics and clusters accurately.

Conclusion Validity: We manually mapped the user questions to relevant model card sections based on our understanding. Two authors mapped the questions independently to reduce bias. Future research should explore more objective methods for mapping user questions to model card sections.

External Validity: We considered one model card template for our analysis, which might introduce bias since other templates could offer different insights. Future research should consider using a range of templates to strengthen and generalize the findings. Additionally, we focus solely on discussions in model repositories on HF. There might be more discussions about these HF models on other platforms like GitHub and Discord, which we did not consider, which could affect the external validity of our study. Moreover, our findings may not generalize to questions asked about models found on other model hubs.

IX. Related Work

A. Question and Answering (Q&A) Communities

Numerous studies have analyzed questions in various software engineering domains, such as web development [6], mobile development [45], game development [25], software security [11] and testing [49], and many more [43], [32], [39], [7], [58], [4]. Roy et al. [46] recently conducted a systematic review of 133 articles on community question-answering sites (CQAs) that use traditional machine learning and deep learning methods to explore the growing volume of CQA content. They identified key research themes in the literature including question quality, answer quality, and expert identification, with popular platforms being Yahoo! Answers, Stack Exchange, and Stack Overflow (SO).

Closest to our work, Yang et al. [59] studied GitHub issues of open-source AI repositories to investigate the problems during the process of employing AI systems to assist developers. They identified 13 categories of problems that developers are likely to encounter in open-source AI repositories. Our study is the first to analyze user questions on ML models and their usage from HF model community discussions, focusing on the effort involved in responding to questions and the topics of those questions.

Supporting and Improving Answering: Several studies have focused on supporting Q&A to improve user experience. Wang et al. [57] made a set of suggestions to information seekers in the MSDN Visual C# General Forum on how to make their questions be answered faster. Wang et al. [55] suggested Q&A website designers improve their incentive systems to motivate non-frequent answerers to be more active and answer questions faster. Wang et al. [56] proposed an automatic tag recommendation system to improve the organization of questions that helps answerers find questions on their topic of interest. Nadi and Treude [35] developed and compared four potential approaches to identify key sentences from SO answers to help users decide whether an answer is relevant to their task and context. Luong et al. [30] developed ArSeek, a context-specific algorithm to capture relevant information from discussions, allowing developers to access useful API resources quickly. Abbasiantaeb et al. [1] used Transformer-based language models to rank candidate answers for each question. Ahmed et al. [3] used CNN-based and BERT-based models to recommend highlighted content with different formatting styles in SO. Lill et al. [28] proposed an approach to automatically identify repeated questions and suggest answers from previous discussions on SO and Discord. In addition, there are many recent studies that focus on domain-specific automated question answering [14], [60], [50], [24]. In our study, we propose the use of model cards to address common user questions to improve question-answering.

B. Studies on the HF Community

There are a few recent studies that have examined the metadata of repositories in HF and its community. Osborrne et al. [38] analyzed various aspects of development activities in HF like interactions in model, dataset, and space repositories; collaboration in model repositories; and model adoption in spaces, to understand collaborative practices in the open AI ecosystem. Castaño et al. [9] examined the status and evolution of the HF community by analyzing changes in popularity, framework usage, tag and dataset trends, and key author groups. Closest to our work, Taraghi et al. [51] conducted an empirical study on the HF discussion forum and identified 17 categories of challenges, while our study focuses on understanding questions asked related to models in the HF model repositories' discussions. Taraghi et al.'s finding of the most prominent category of challenges, Model Usage & Understanding, and Training Pipeline, supports our proposal to add more development-related content on training and a section on how to use the model in the model card.

C. Model Documentation Analysis and Improvement

Current studies examine the content of existing model documentation to identify gaps and scope of improvements in model documentation [8], [27], [41], [37]. Bhat et al. [8] and Liang et al. [27] showed that model cards do not often contain enough information about different sections of the proposed standard template. Pepe et al. [41] highlighted the need for better documentation of training datasets, biases, and licenses in pre-trained models to improve transparency and mitigate potential biases and legal issues. Oreamuno et al. [37] found that many models and datasets in the HF store lack comprehensive documentation, either failing to meet user needs or lacking enforcement. The study demonstrated inconsistencies in ethics and transparency-related documentation for ML models and datasets, indicating the need for improved practices to address ethical concerns, biases, and limitations. Additionally, they suggested adding categories for model versioning and attribution in documentation standards.

Crisan et al. [10] introduced a new design for model documentation, the Interactive Model Card (IMC), to make it more understandable and actionable for a wider range of stakeholders, by incorporating interactive elements and human-centered design principles into the model documentation process. The authors provided guidelines for IMC based on the feedback from both AI experts and non-expert analysts which enhance the transparency and interpretability of model details for a diverse range of stakeholders. Tsay et al. [53] identified challenges in documenting AI models, including the reliance on manual processes and lack of standardized practices, which lead to inconsistencies and information gaps. As a solution, they extracted metadata relevant to model documentation from software repositories and created a searchable catalog using this metadata to improve model documentation.

X. Conclusion

In this paper, we analyze 11,278 discussions about ML models on HF. We find that while answering questions does not require much effort, questions that do not receive a response in 24 hours are unlikely to be answered at all. Also, 71.9% of the question-containing discussions are open, of which, 40.1% are open without a response. A topic analysis of the question-containing posts shows that most discussions are related to technical details of model development and model usage, and requests for model variants. Through a manual analysis of the questions, we found that 42.5% of the questions could be answered following a standard model card template, and adding (sub)sections on model usage and hardware requirements would cover even more frequently asked questions. With our suggestion for model card enhancements, we believe support efforts for model developers can be simplified. The suggested improved model documentation can optimize the question-answer process and decrease wait time for an answer.

Acknowledgements

The work described in this paper has been supported by the ECE-Huawei Research Initiative (HERI) at the University of Alberta.

References

[1] Z. Abbasiantaeb and S. Momtazi, "Entity-aware answer sentence selection for question answering with transformer-based language models," *Journal of Intelligent Information Systems*, vol. 59, no. 3, pp. 755–777, 2022.

[2] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia Computer Science*, vol. 189, pp. 191–194, 2021, aI in Computational Linguistics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921012199

[3] S. S. Ahmed, S. Wang, Y. Tian, T.-H. P. Chen, and H. Zhang, "Studying and recommending information highlighting in Stack Overflow answers," *Information and Software Technology*, vol. 172, p. 107478, 2024.

[4] L. Albesher, R. Aldossari, and R. Alfayez, "An Observational Study on React Native (RN) Questions on Stack Overflow (SO)," *IET Software*, vol. 2023, no. 1, p. 6613434, 2023.

[5] M. T. Amith, L. Cui, D. Zhi, K. Roberts, X. Jiang, F. Li, E. Yu, and C. Tao, "Toward a standard formal semantic representation of the model card report," *BMC bioinformatics*, vol. 23, no. Suppl 6, p. 281, 2022.

[6] K. Bajaj, K. Pattabiraman, and A. Mesbah, "Mining Questions Asked by Web Developers," in *Proceedings of the 11th Working Conference on Mining Software Repositories*, ser. MSR 2014. New York, NY, USA: Association for Computing Machinery, 2014, p. 112–121. [Online]. Available: https://doi.org/10.1145/2597073.2597083

[7] M. Begoug, N. Bessghaier, A. Ouni, E. A. AlOmar, and M. W. Mkaouer, "What Do Infrastructure-as-Code Practitioners Discuss: An Empirical Study on Stack Overflow," in *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2023, pp. 1–12.

[8] A. Bhat, A. Coursey, G. Hu, S. Li, N. Nahar, S. Zhou, C. Kästner, and J. L. Guo, "Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.

[9] J. Castaño, S. Martínez-Fernández, X. Franch, and J. Bogner, "Analyzing the Evolution and Maintenance of ML Models on Hugging Face," in *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*. IEEE, 2024, pp. 607–618.

[10] A. Crisan, M. Drouhard, J. Vig, and N. Rajani, "Interactive Model Cards: A Human-Centered Approach to Model Documentation," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 427–439. [Online]. Available: https://doi.org/10.1145/3531146.3533108

[11] R. Croft, Y. Xie, M. Zahedi, M. A. Babar, and C. Treude, "An empirical study of developers' discussions about security challenges of different programming languages," *Empirical Software Engineering*, vol. 27, pp. 1–52, 2022.

[12] K. DENECKE, "What Do Autistic People Discuss on Twitter? An Approach Using BERTopic Modelling," *Caring is Sharing—Exploiting the Value in Data for Health and Innovation: Proceedings of MIE 2023*, vol. 302, p. 403, 2023.

[13] B. Ding, C. Qin, L. Liu, Y. K. Chia, S. Joty, B. Li, and L. Bing, "Is GPT-3 a Good Data Annotator?" *arXiv preprint arXiv:2212.10450*, 2022.

[14] L. do Nascimento Vale and M. de Almeida Maia, "Towards a question answering assistant for software development using a transformer-based language model," in *2021 IEEE/ACM Third International Workshop on Bots in Software Engineering (BotSE)*. IEEE, 2021, pp. 39–42.

[15] A. Donald, A. Galanopoulos, E. Curry, E. Muñoz, I. Ullah, M. Waskow, M. Dabrowski, and M. Kalra, "Towards a Semantic Approach for Linked Dataspace, Model and Data Cards," in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1468–1473.

[16] M. A. dos Santos, M. J. Sánchez-Franco, F. C. Moreno, and M. H. G. Serrano, "Modelling the structure of the sports management research field using the BERTopic approach," *Retos: nuevas tendencias en educación física, deporte y recreación*, no. 47, pp. 648–663, 2023.

[17] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498

[18] L. Gan, T. Yang, Y. Huang, B. Yang, Y. Y. Luo, L. W. C. Richard, and D. Guo, "Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic," in *International Symposium on Artificial Intelligence and Robotics*, H. Lu and J. Cai, Eds., Springer. Singapore: Springer Nature Singapore, 2024, pp. 376–391.

[19] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-Trained Models: Past, Present and Future," *AI Open*, vol. 2, pp. 225–250, 2021.

[20] G. Hristova and N. Netov, "Media Coverage and Public Perception of Distance Learning During the COVID-19 Pandemic: A Topic Modeling Approach Based on BERTopic," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 2259–2264.

[21] N. Jagdishbhai and K. Y. Thakkar, "Exploring the capabilities and limitations of GPT and Chat GPT in natural language processing," *Journal of Management Research and Analysis*, vol. 10, no. 1, pp. 18–20, 2023.

[22] S. M. Jain, "Hugging Face," in *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. Springer, 2022, pp. 51–67. [Online]. Available: https://doi.org/10.1007/978-1-4842-8844-3_4

[23] W. Jiang, J. 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 *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*. IEEE, 2024, pp. 431–443.

[24] S. Kabir, D. N. Udo-Imeh, B. Kou, and T. Zhang, "Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–17.

[25] A. Kamienski and C.-P. Bezemer, "An empirical study of Q&A websites for game developers," *Empirical Software Engineering*, vol. 26, no. 6, p. 115, 2021.

[26] A. Kathikar, A. Nair, B. Lazarine, A. Sachdeva, and S. Samtani, "Assessing the Vulnerabilities of the Open-Source Artificial Intelligence (AI) Landscape: A Large-Scale Analysis of the Hugging Face Platform," in *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2023, pp. 1–6.

[27] W. Liang, N. Rajani, X. Yang, E. Ozoani, E. Wu, Y. Chen, D. S. Smith, and J. Zou, "What's documented in AI? Systematic Analysis of 32K AI Model Cards," *arXiv preprint arXiv:2402.05160*, 2024.

[28] A. Lill, A. N. Meyer, and T. Fritz, "On the Helpfulness of Answering Developer Questions on Discord with Similar Conversations and Posts from the Past," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 2024, pp. 1–13.

[29] Luca Papariello, "xlm-roberta-base-language-detection (revision 9865598)," 2024. [Online]. Available: https://huggingface.co/papluca/xlm-roberta-base-language-detection

[30] K. Luong, M. Hadi, F. Thung, F. Fard, and D. Lo, "ARSeek: Identifying API Resource using Code and Discussion on Stack Overflow," in *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, 2022, pp. 331–342.

[31] Maarten Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[32] K. Mahmood, G. Rasool, F. Sabir, and A. Athar, "An Empirical Study of Web Services Topics in Web Developer Discussions on Stack Overflow," *IEEE Access*, vol. 11, pp. 9627–9655, 2023.

[33] P. E. McKnight and J. Najab, "Mann–Whitney U Test," *The SAGE Encyclopedia of Research Design*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:118856424

[34] 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.

[35] S. Nadi and C. Treude, "Essential Sentences for Navigating Stack Overflow Answers," in *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2020, pp. 229–239.

[36] J. L. Nunes, G. D. Barbosa, C. S. de Souza, and S. D. Barbosa, "Using Model Cards for ethical reflection on machine learning models: an interview-based study," *Journal on Interactive Systems*, vol. 15, no. 1, pp. 1–19, 2024.

[37] E. L. Oreamuno, R. F. Khan, A. A. Bangash, C. Stinson, and B. Adams, "The State of Documentation Practices of Third-party Machine Learning Models and Datasets," *IEEE Software*, 2024.

[38] C. Osborne, J. Ding, and H. R. Kirk, "The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub," *Journal of Computational Social Science*, pp. 1–39, 2024.

[39] A. Ouni, I. Saidani, E. Alomar, and M. W. Mkaouer, "An Empirical Study on Continuous Integration Trends, Topics and Challenges in Stack Overflow," in *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, 2023, pp. 141–151.

[40] F. Pascual. (2022) Introducing the Private Hub: A New Way to Build With Machine Learning. [Accessed 05-06-2024]. [Online]. Available: https://huggingface.co/blog/introducing-private-hub

[41] F. Pepe, V. Nardone, A. Mastropaolo, G. Bavota, G. Canfora, and M. Di Penta, "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*, 2024, pp. 370–381.

[42] S. V. Raju, B. K. Bolla, D. K. Nayak, and J. Kh, "Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*. IEEE, 2022, pp. 1–6.

[43] P. Rani, M. Birrer, S. Panichella, M. Ghafari, and O. Nierstrasz, "What Do Developers Discuss about Code Comments?" in *2021 IEEE 21st International Working Conference on Source Code Analysis and Manipulation (SCAM)*. IEEE, 2021, pp. 153–164.

[44] J. Romano, J. D. Kromrey, J. Coraggio, J. Skowronek, and L. Devine, "Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate

choices," in *Annual meeting of the Southern Association for Institutional Research*. Citeseer, 2006, pp. 1–51.

[45] C. Rosen and E. Shihab, "What are mobile developers asking about? a large scale study using stack overflow," *Empirical Software Engineering*, vol. 21, pp. 1192–1223, 2016.

[46] P. K. Roy, S. Saumya, J. P. Singh, S. Banerjee, and A. Gutub, "Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 95–117, 2023.

[47] S. Samsir, R. S. Saragih, S. Subagio, R. Aditiya, and R. Watrianthos, "BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, pp. 1514–1520, 2023.

[48] I. Scarpino, C. Zucco, R. Vallelunga, F. Luzza, and M. Cannataro, "Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration," *BioTech*, vol. 11, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2673-6284/11/3/41

[49] M. Swillus and A. Zaidman, "Sentiment overflow in the testing stack: Analyzing software testing posts on Stack Overflow," *Journal of Systems and Software*, vol. 205, p. 111804, 2023.

[50] L. Szerszen, "Question answering on introductory Java programming concepts using the Transformer," 2021.

[51] M. Taraghi, G. Dorcelus, A. Foundjem, F. Tambon, and F. Khomh, "Deep Learning Model Reuse in the HuggingFace Community: Challenges, Benefit and Trends," *arXiv preprint arXiv:2401.13177*, 2024.

[52] T. R. Toma, B. Grewal, and C.-P. Bezemer, "Replication Package," 2024. [Online]. Available: https://github.com/asgaardlab/hf-question-answer

[53] J. Tsay, A. Braz, M. Hirzel, A. Shinnar, and T. Mummert, "AIMMX: Artificial Intelligence Model Metadata Extractor," in *Proceedings of the 17th International Conference on Mining Software Repositories*, 2020, pp. 81–92.

[54] M. Uncovska, B. Freitag, S. Meister, and L. Fehring, "Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany," *NPJ Digital Medicine*, vol. 6, no. 1, p. 115, 2023.

[55] S. Wang, T.-H. Chen, and A. E. Hassan, "Understanding the factors for fast answers in technical Q&A websites: An empirical study of four stack exchange websites," *Empirical Software Engineering*, vol. 23, pp. 1552–1593, 2018.

[56] S. Wang, D. Lo, B. Vasilescu, and A. Serebrenik, "ENTAGREC++: An enhanced tag recommendation system for software information sites," *Empirical Software Engineering*, vol. 23, pp. 800–832, 2018.

[57] Y. Wang, "Making your programming questions be answered quickly: A content oriented study to technical Q&A forum," in *10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*. IEEE, 2014, pp. 368–377.

[58] Z. Wang, T.-H. P. Chen, H. Zhang, and S. Wang, "An empirical study on the challenges that developers encounter when developing Apache Spark applications," *Journal of Systems and Software*, vol. 194, p. 111488, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121222001674

[59] Z. Yang, C. Wang, J. Shi, T. Hoang, P. Kochhar, Q. Lu, Z. Xing, and D. Lo, "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)*. IEEE, 2023, pp. 79–91.

[60] H. Zhu, P. Tiwari, A. Ghoneim, and M. S. Hossain, "A Collaborative AI-Enabled Pretrained Language Model for AIoT Domain Question Answering," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3387–3396, 2021.