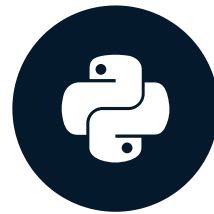


# Introduction to Hugging Face

WORKING WITH HUGGING FACE



**Jacob H. Marquez**  
Lead Data Engineer

# What is Hugging Face?



- Collaboration platform
- Open-source machine learning
- Text, vision, and audio tasks
- Models, datasets, frameworks
- Reduce barriers to entry

<sup>1</sup> <https://huggingface.co/>

# In this course

- Navigate and use the Hugging Face Hub
- Explore models and datasets
- Build pipelines for text, image, and audio data
- Fine-tuning, generation, embeddings, and semantic search



# Large Language Models

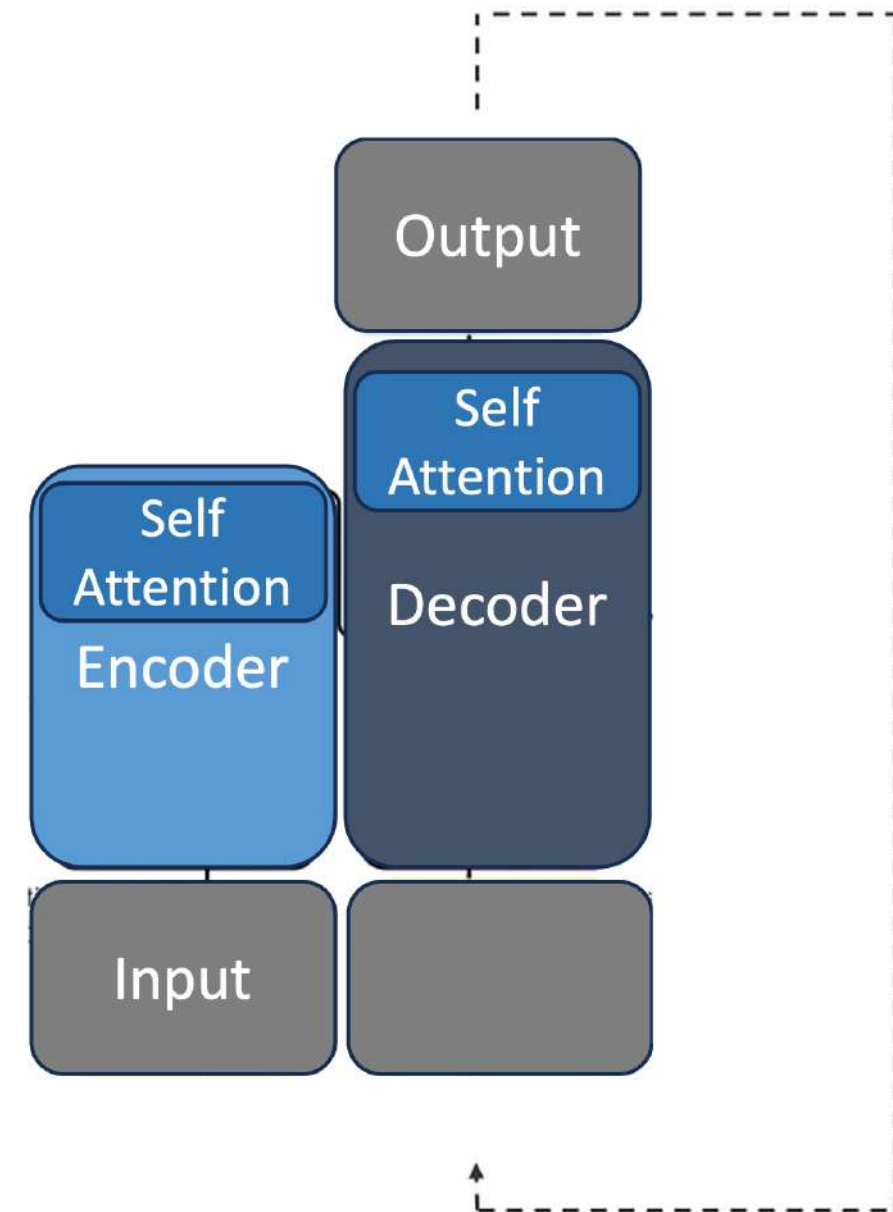
- LLMs
- Understand and generate human-like text
- Massive amounts of data
- Learn patterns in sequences



<sup>1</sup> [https://en.wikipedia.org/wiki/Large\\_language\\_model](https://en.wikipedia.org/wiki/Large_language_model)

# Large Language Models

- LLMs
- Understand and generate human-like text
- Massive amounts of data
- Learn patterns in sequences
- Transformer architecture



<sup>1</sup> <https://towardsdatascience.com/transformers-89034557de14>

# Large Language Models

- LLMs
- Understand and generate human-like text
- Massive amounts of data
- Learn patterns in sequences
- Transformer architecture
- Popular options are GPT and Llama

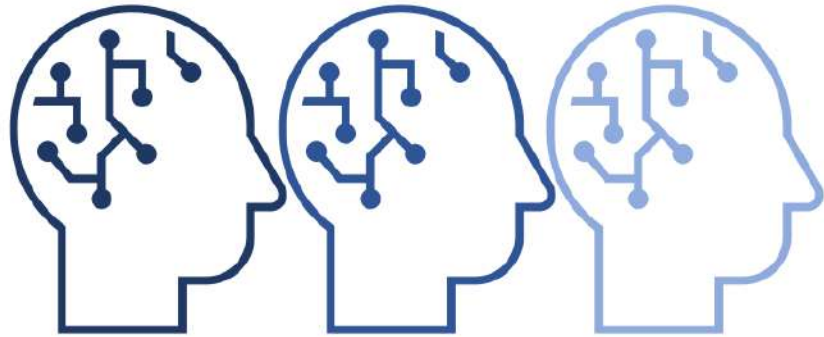


**ChatGPT**

**LLaMA**  
by  **Meta**



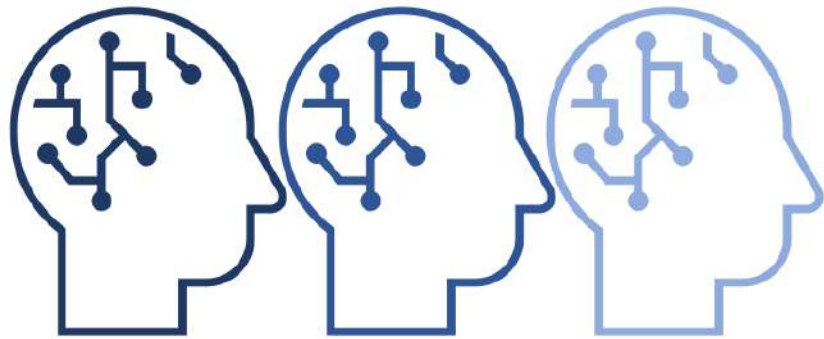
# Benefits of Hugging Face



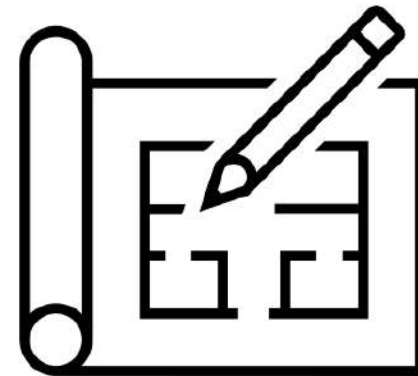
## Access to Models

- Faster experimentation

# Benefits of Hugging Face



## Access to Models

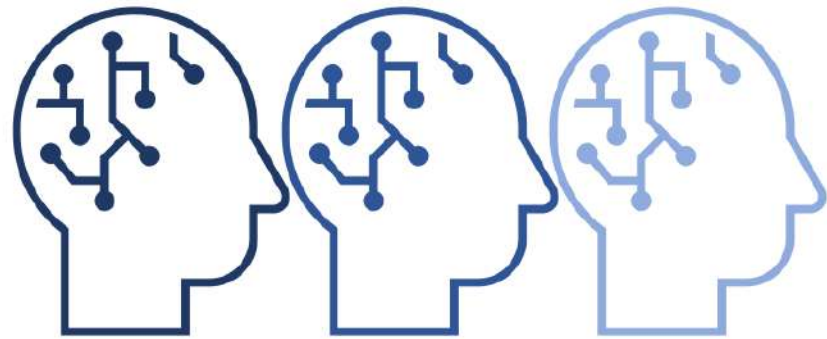


## Frameworks

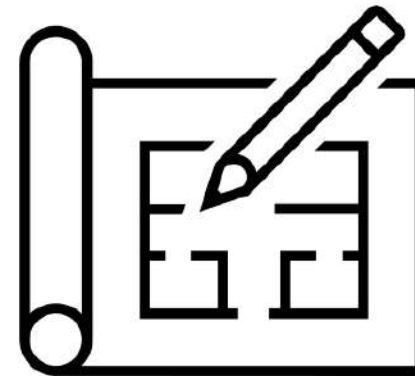
- Faster experimentation
- Supports every step of the process



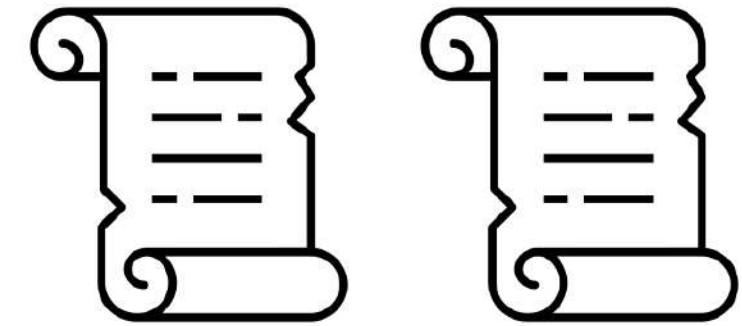
# Benefits of Hugging Face



## Access to Models



## Frameworks



## Documentation

- Faster experimentation
- Supports every step of the process
- Smoother adoption

# Deciding when to use

## Use Hugging Face

- Quick way to use ML tasks
- Don't have deep ML expertise
- Testing several models
- Dataset needed

## Use another solution

- Slow computer
- Highly customized architectures
- Domain specific needs not yet met
- Not leveraging advanced ML techniques

# Installing Hugging Face

Hugging Face

```
pip install transformers datasets
```

ML Framework

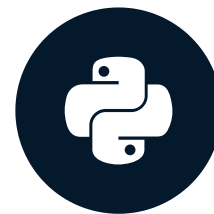
```
pip install torch torchvision torchaudio
```

<sup>1</sup> <https://pytorch.org/>

**Let's practice!**  
WORKING WITH HUGGING FACE

# Transformers and the Hub

WORKING WITH HUGGING FACE



**Jacob H. Marquez**  
Lead Data Engineer

# Transformers - the Hugging Face package



## Transformers

build **passing** license Apache-2.0 website **online** release v4.37.2 Contributor Covenant v2.0 adopted DOI 10.5281/zenodo.7391177

English | 简体中文 | 繁體中文 | 한국어 | Español | 日本語 | हिन्दी | Русский | Português | తెలుగు | Français | Deutsch |

State-of-the-art Machine Learning for JAX, PyTorch and TensorFlow



### Part of the Hugging Face course!

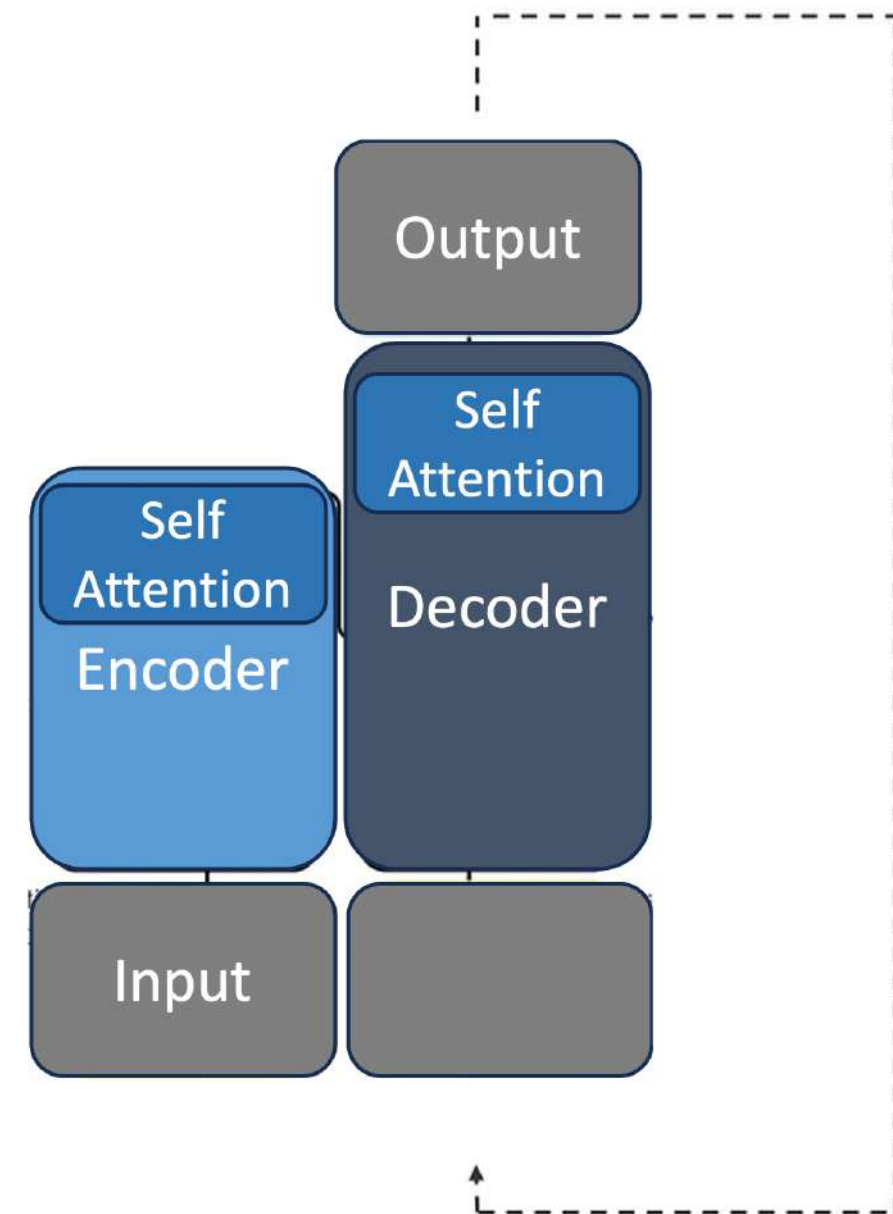
🧑🎓 Transformers provides thousands of pretrained models to perform tasks on different modalities such as text,

<sup>1</sup> <https://github.com/huggingface/transformers>



# Transformers - the model architecture

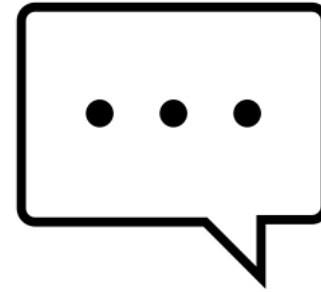
- Neural network models
- Learn context and understanding
- Core components:
  - Encoder
  - Decoder
  - Self-attention mechanism
- Transform input to numerical representations
- Helps model understand context of the input



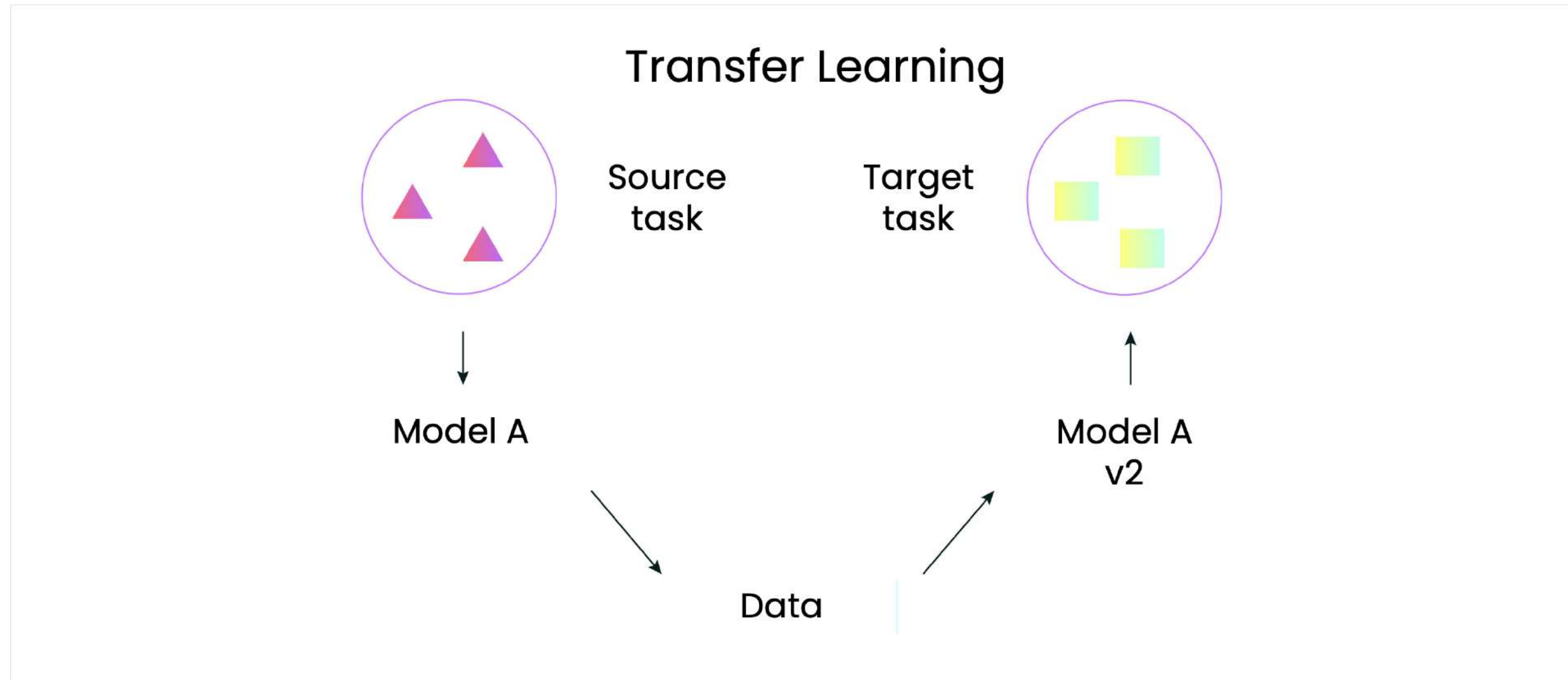
<sup>1</sup> <https://www.turing.com/kb/brief-introduction-to-transformers-and-their-power>

# Uses cases of transformers

- Use cases for text, image, and vision
- Classification for all three
- Automatic speech recognition
- Text summarization
- Object detection for autonomous driving



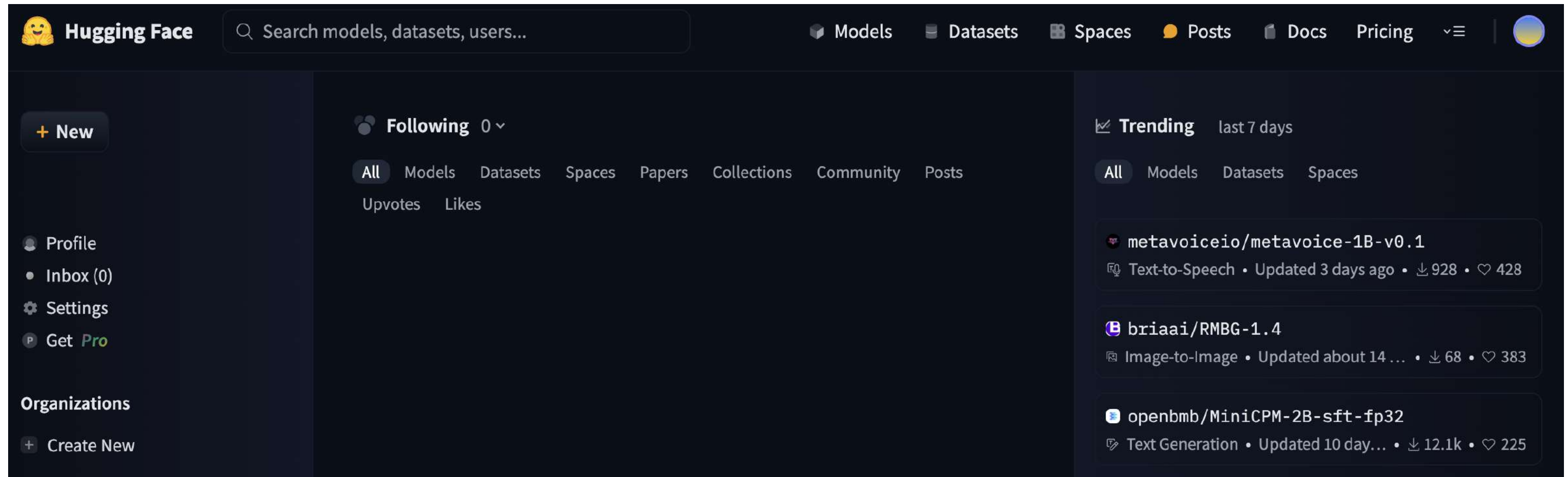
# A key benefit of transformers



- Enables Hugging Face models to perform well on new tasks with little data

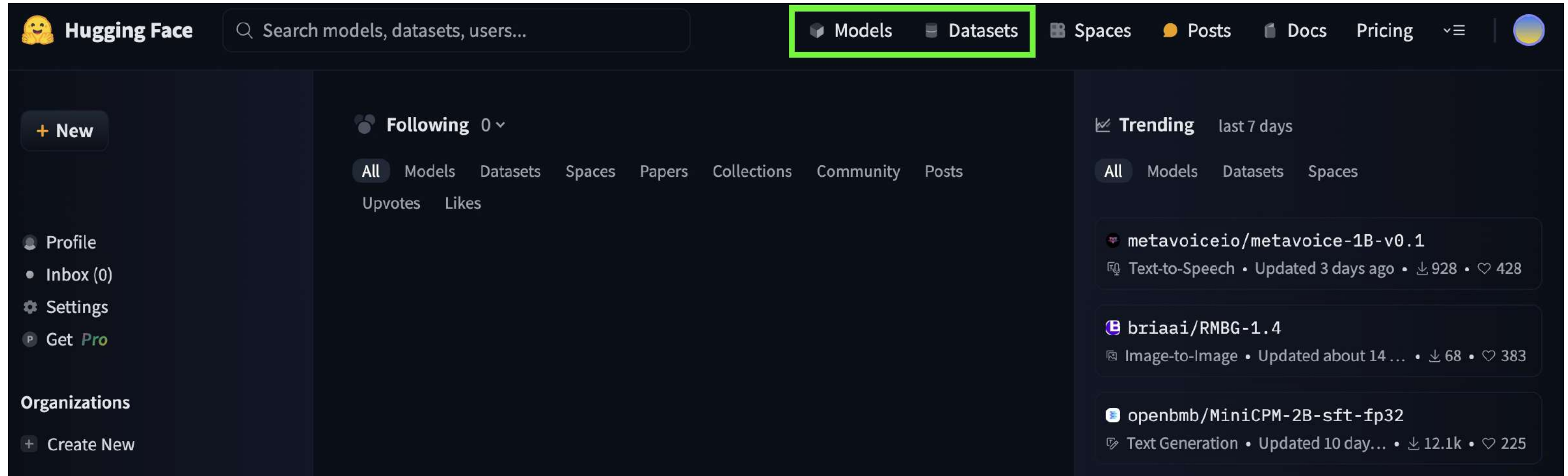
<sup>1</sup> <https://www.topbots.com/transfer-learning-in-nlp/#transfer-learning>

# The Hub



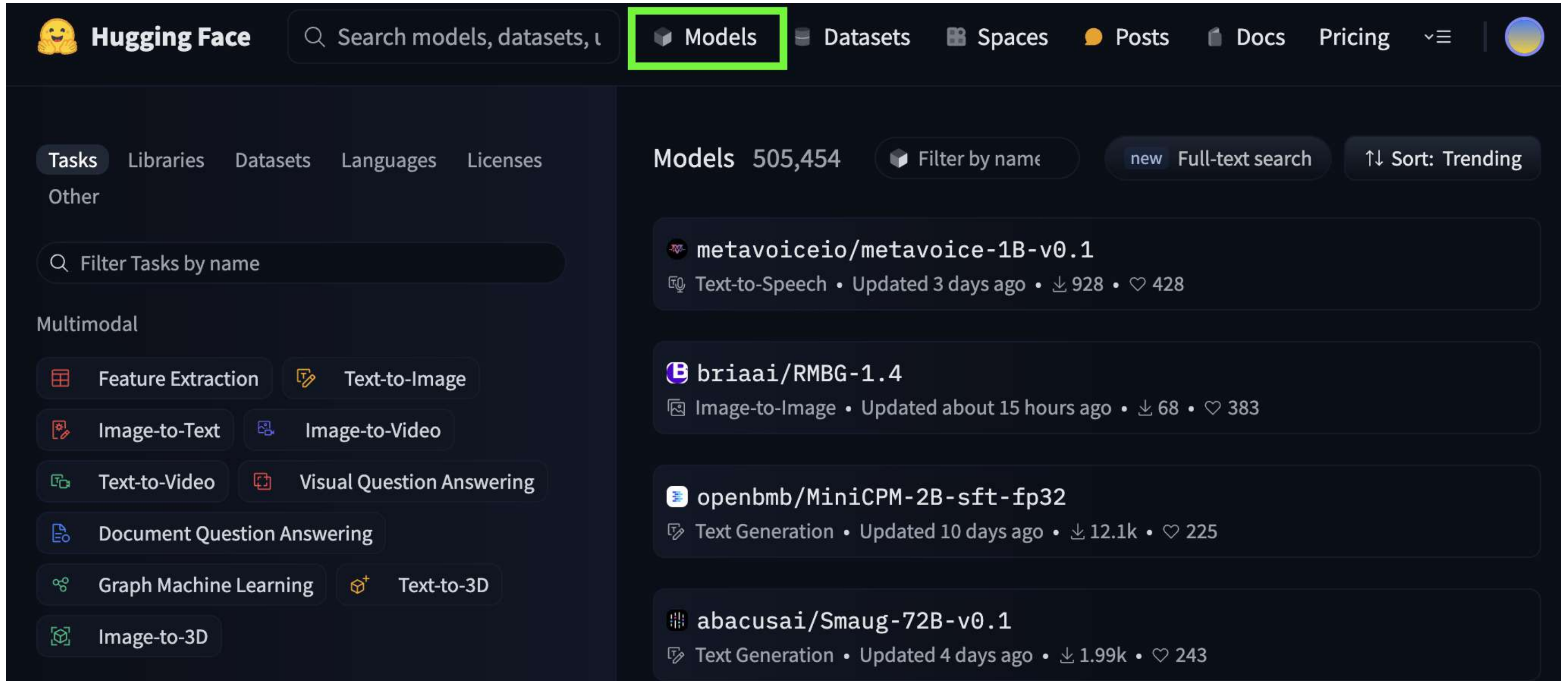
<sup>1</sup> <https://huggingface.co/>

# Navigating the Hub



<sup>1</sup> <https://huggingface.co/>

# Searching for models



The screenshot shows the Hugging Face website's 'Models' page. The 'Models' tab is highlighted with a green box in the top navigation bar. The left sidebar contains filters for Tasks, Libraries, Datasets, Languages, Licenses, and Other, with a search bar for 'Filter Tasks by name'. Below this, a 'Multimodal' section lists various task categories like Feature Extraction, Text-to-Image, Image-to-Text, Image-to-Video, Text-to-Video, Visual Question Answering, Document Question Answering, Graph Machine Learning, Text-to-3D, and Image-to-3D. The main content area displays a list of models, starting with 'metavoiceio/metavoice-1B-v0.1' (Text-to-Speech), 'briaai/RMBG-1.4' (Image-to-Image), 'openbmb/MiniCPM-2B-sft-fp32' (Text Generation), and 'abacusai/Smaug-72B-v0.1' (Text Generation). Each model entry includes its name, task type, update time, download count, and heart count.

**Hugging Face** Search models, datasets, l **Models** Datasets Spaces Posts Docs Pricing

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Image-to-Video
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning
- Text-to-3D
- Image-to-3D

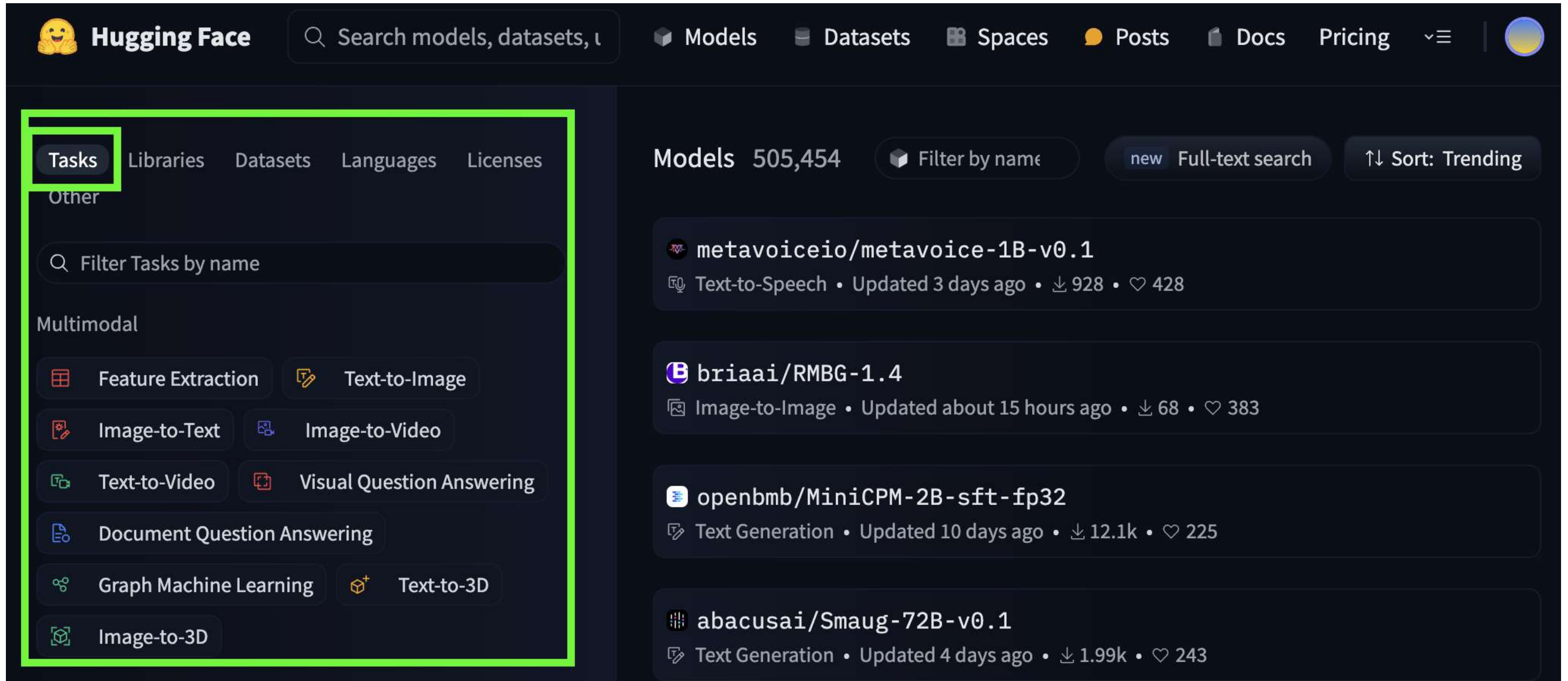
**Models** 505,454 Filter by name new Full-text search Sort: Trending

- metavoiceio/metavoice-1B-v0.1**  
Text-to-Speech • Updated 3 days ago • 928 • 428
- briaai/RMBG-1.4**  
Image-to-Image • Updated about 15 hours ago • 68 • 383
- openbmb/MiniCPM-2B-sft-fp32**  
Text Generation • Updated 10 days ago • 12.1k • 225
- abacusai/Smaug-72B-v0.1**  
Text Generation • Updated 4 days ago • 1.99k • 243

<sup>1</sup> <https://huggingface.co/models>



# Searching for models



The screenshot shows the Hugging Face website interface. The top navigation bar includes the Hugging Face logo, a search bar, and links to Models, Datasets, Spaces, Posts, Docs, and Pricing. The left sidebar is highlighted with a green border and contains the following sections:

- Tasks** (highlighted with a green box)
- Libraries
- Datasets
- Languages
- Licenses
- Other

Below the sidebar, there is a search bar labeled "Filter Tasks by name". The main content area displays a list of models under the "Models" tab, which shows 505,454 models. The list includes:

- metavoiceio/metavoice-1B-v0.1**: Text-to-Speech • Updated 3 days ago • 928 downloads • 428 likes
- briaai/RMBG-1.4**: Image-to-Image • Updated about 15 hours ago • 68 downloads • 383 likes
- openbmb/MiniCPM-2B-sft-fp32**: Text Generation • Updated 10 days ago • 12.1k downloads • 225 likes
- abacusai/Smaug-72B-v0.1**: Text Generation • Updated 4 days ago • 1.99k downloads • 243 likes

<sup>1</sup> <https://huggingface.co/models>

# Searching for models

The screenshot shows the Hugging Face website interface. In the top navigation bar, the 'Hugging Face' logo is on the left, followed by a search bar and links for 'Models', 'Datasets', 'Spaces', 'Posts', 'Docs', and 'Pricing'. Below the navigation bar, the left sidebar contains a list of categories: 'Tasks', 'Libraries', 'Datasets', 'Languages' (highlighted with a green box), and 'Licenses'. Under 'Languages', there is a search bar 'Filter Languages by name' and a grid of language buttons including English, Chinese, French, German, Spanish, Japanese, Korean, Russian, Italian, Portuguese, Arabic, Hindi, Turkish, Dutch, Swedish, multilingual, Polish, Indonesian, Vietnamese, Finnish, Enawené-Nawé, Romanian, Thai, Ukrainian, and Persian. The main content area is titled 'Models 505,454' and includes a 'Filter by name' button, a 'new Full-text search' button, and a 'Sort: Trending' button. Below this, a list of models is displayed, each with its name, task type, update time, and download/like counts. The models listed are: 'metavoiceio/metavoice-1B-v0.1' (Text-to-Speech, Updated 3 days ago, 928 downloads, 428 likes), 'briaai/RMBG-1.4' (Image-to-Image, Updated about 15 hours ago, 68 downloads, 383 likes), 'openbmb/MiniCPM-2B-sft-fp32' (Text Generation, Updated 10 days ago, 12.1k downloads, 225 likes), and 'abacusai/Smaug-72B-v0.1' (Text Generation, Updated 4 days ago, 1.99k downloads, 243 likes).

<sup>1</sup> <https://huggingface.co/models>

# Searching for models

The screenshot shows the Hugging Face website interface. The top navigation bar includes the Hugging Face logo, a search bar, and links to Models, Datasets, Spaces, Posts, Docs, and Pricing. The left sidebar contains a list of categories: Tasks, Libraries (highlighted with a green box), Datasets, Languages, Licenses, and Other. Below the sidebar, there is a search bar for filtering libraries by name and a grid of library tags such as PyTorch, TensorFlow, JAX, Transformers, TensorBoard, Safetensors, Diffusers, PEFT, Stable-Baselines3, ONNX, Unity ML-Agents, GGUF, Sentence Transformers, Keras, Timm, Flair, Sample Factory, SetFit, Adapters, Transformers.js, spaCy, ESPnet, fastai, Core ML, and NeMo. The main content area displays a list of models with the following details:

- Models** 505,454
- Filter by name**
- new Full-text search**
- Sort: Trending**
- metavoiceio/metavoice-1B-v0.1**  
Text-to-Speech • Updated 3 days ago • 928 • 428
- briaai/RMBG-1.4**  
Image-to-Image • Updated about 15 hours ago • 68 • 383
- openbmb/MiniCPM-2B-sft-fp32**  
Text Generation • Updated 10 days ago • 12.1k • 225
- abacusai/Smaug-72B-v0.1**  
Text Generation • Updated 4 days ago • 1.99k • 243

<sup>1</sup> <https://huggingface.co/models>



# Model cards

The screenshot shows the Hugging Face interface for the `openai/whisper-large-v3` model. The top navigation bar includes the Hugging Face logo, a search bar, and links to Models, Datasets, Spaces, Posts, Docs, and Pricing. The model card header displays the model name, a 'like' button with 1.63k likes, and various tags: Automatic Speech Recognition, Transformers, PyTorch, JAX, Safetensors, 99 languages, whisper, and audio. Below the tags are links to the hf-asr-leaderboard, Inference Endpoints, arXiv papers (2212.04356 and 2311.00430), and the license (apache-2.0). The main content area has tabs for Model card, Files, and Community (80 members). The 'Model card' tab is active, showing the title 'Whisper' and a description: 'Whisper is a pre-trained model for automatic speech recognition (ASR) and speech translation. Trained on 680k hours of labelled data, Whisper models demonstrate a strong ability to generalise to'. On the right side of the card, there is a section for 'Downloads last month' showing 931,146 downloads with a line graph. Below this, the 'Safetensors' section shows the model size as 1.54B params and the tensor type as FP16.

**Hugging Face** Search models, datasets, u Models Datasets Spaces Posts Docs Pricing

**openai/whisper-large-v3** like 1.63k

Automatic Speech Recognition Transformers PyTorch JAX Safetensors 99 languages whisper audio

hf-asr-leaderboard Inference Endpoints arxiv:2212.04356 arxiv:2311.00430 License: apache-2.0

Model card Files Community 80

Edit model card

## Whisper

Whisper is a pre-trained model for automatic speech recognition (ASR) and speech translation. Trained on 680k hours of labelled data, Whisper models demonstrate a strong ability to generalise to

Downloads last month  
**931,146**

**Safetensors** Model size 1.54B params

Tensor type FP16

<sup>1</sup> <https://huggingface.co/openai/whisper-large-v3>

# Using huggingface\_hub

```
pip install huggingface_hub
```

```
from huggingface_hub import HfApi  
api = HfApi()  
list(api.list_models())
```

```
[ModelInfo: {  
  {'_id': '622fea36174feb5439c2e4be',  
    'author': 'cardiffnlp',  
    ...}]
```

<sup>1</sup> [https://github.com/huggingface/huggingface\\_hub](https://github.com/huggingface/huggingface_hub)

# Using huggingface\_hub

```
models = api.list_models(  
    filter=ModelFilter(  
        task="text-classification"),  
        sort="downloads",  
        direction=-1,  
        limit=5  
    )  
)  
  
modelList = list(models)  
  
print(modelList[0])
```

```
Model Name: albert/albert-base-v1, Tags: [...]
```

- `task` searches for specified task
- `sort` will order the list
- `direction` provides the direction of the sorted order
  - -1 for descending
  - all other numbers for ascending
- `limit` will limit the number of models returned

<sup>1</sup> [https://github.com/huggingface/huggingface\\_hub](https://github.com/huggingface/huggingface_hub)



# Saving a model locally

```
# Import AutoModel
from transformers import AutoModel

modelId = "distilbert-base-uncased-finetuned-sst-2-english"

# Download model using the modelId
model = AutoModel.from_pretrained(modelId)

# Save the model to a local directory
model.save_pretrained(save_directory=f"models/{modelId}")
```

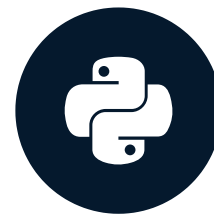
- Be mindful of storage!

<sup>1</sup> [https://huggingface.co/docs/transformers/model\\_doc/auto#transformers.AutoModel](https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoModel)

**Let's practice!**  
WORKING WITH HUGGING FACE

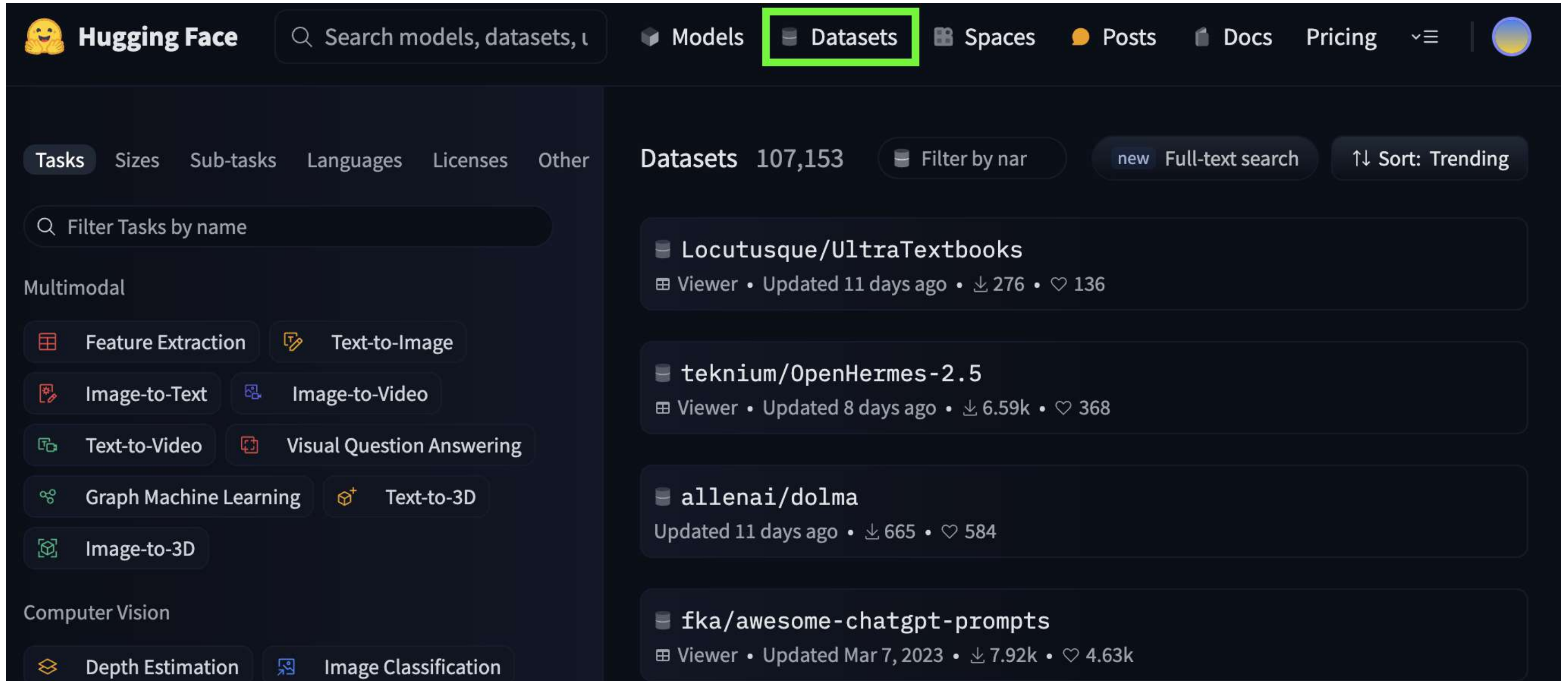
# Working with datasets

WORKING WITH HUGGING FACE



**Jacob H. Marquez**  
Lead Data Engineer

# Datasets in Hugging Face



The screenshot shows the Hugging Face website's 'Datasets' page. The top navigation bar includes the Hugging Face logo, a search bar, and links to Models, Datasets (highlighted with a green box), Spaces, Posts, Docs, and Pricing. The left sidebar contains filters for Tasks, Sizes, Sub-tasks, Languages, Licenses, and Other, along with a search bar for tasks and a list of task categories like Multimodal, Feature Extraction, Text-to-Image, Image-to-Text, Image-to-Video, Text-to-Video, Visual Question Answering, Graph Machine Learning, Text-to-3D, Image-to-3D, Computer Vision, Depth Estimation, and Image Classification. The main content area displays a list of datasets with the following details:

- Datasets 107,153** (Filter by nar, new Full-text search, Sort: Trending)
- Locutusque/UltraTextbooks**
  - Viewer • Updated 11 days ago • 276 • 136
- teknium/OpenHermes-2.5**
  - Viewer • Updated 8 days ago • 6.59k • 368
- allenai/dolma**
  - Updated 11 days ago • 665 • 584
- fka/awesome-chatgpt-prompts**
  - Viewer • Updated Mar 7, 2023 • 7.92k • 4.63k

<sup>1</sup> <https://huggingface.co/datasets>

# Searching for datasets

The screenshot shows the Hugging Face website interface. At the top, the Hugging Face logo is on the left, followed by a search bar with the placeholder text "Search models, datasets, ...". To the right of the search bar are navigation links: "Models", "Datasets", "Spaces", "Posts", "Docs", and "Pricing". A user profile icon is on the far right.

Below the navigation bar, there is a horizontal menu with tabs: "Tasks", "Sizes", "Sub-tasks", "Languages", "Licenses", and "Other". The "Tasks" tab is highlighted with a green border.

Under the "Tasks" tab, there is a search bar labeled "Filter Tasks by name". Below this, tasks are categorized into "Multimodal" and "Computer Vision".

**Multimodal tasks:**

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Image-to-Video
- Text-to-Video
- Visual Question Answering
- Graph Machine Learning
- Text-to-3D
- Image-to-3D

**Computer Vision tasks:**

- Depth Estimation
- Image Classification

On the right side of the page, the "Datasets" section is displayed. It shows a total of 107,153 datasets. There are filters for "Filter by nar" and "new Full-text search". The sorting is set to "Sort: Trending".

The following datasets are listed:

- Locutusque/UltraTextbooks**  
Viewer • Updated 11 days ago • ⬇ 276 • ❤ 136
- teknium/OpenHermes-2.5**  
Viewer • Updated 8 days ago • ⬇ 6.59k • ❤ 368
- allenai/dolma**  
Updated 11 days ago • ⬇ 665 • ❤ 584
- fka/awesome-chatgpt-prompts**  
Viewer • Updated Mar 7, 2023 • ⬇ 7.92k • ❤ 4.63k

<sup>1</sup> <https://huggingface.co/datasets>

# Dataset cards

The screenshot shows the Hugging Face website interface for the 'imdb' dataset. At the top is the Hugging Face logo and a search bar. Below the search bar are navigation links for Models, Datasets, Spaces, Posts, Docs, and Pricing. The main section displays the 'imdb' dataset card with various filters and tabs. The 'Dataset card' tab is selected, showing a 'Dataset Viewer' section with a split dropdown set to 'train' (25k rows) and a search bar. To the right of the viewer, there are buttons for 'Auto-converted to Parquet', 'API', and 'View in Dataset Viewer'. Further right, it shows 'Downloads last month' as 273,080, and buttons for 'Use in Datasets library' and 'Edit dataset card'.

**Hugging Face** Search models, datasets, ... Models Datasets Spaces Posts Docs Pricing

**Datasets: imdb** like 160

Tasks: Text Classification Sub-tasks: sentiment-classification Languages: English Multilinguality: monolingual

Size Categories: 10K<n<100K Language Creators: expert-generated Annotations Creators: expert-generated Source Datasets: original

Tags: Croissant License: other

Dataset card Viewer Files Community 6

**Dataset Viewer** Auto-converted to Parquet API View in Dataset Viewer

Split  
train · 25k rows

Search this dataset

Downloads last month 273,080

Use in Datasets library

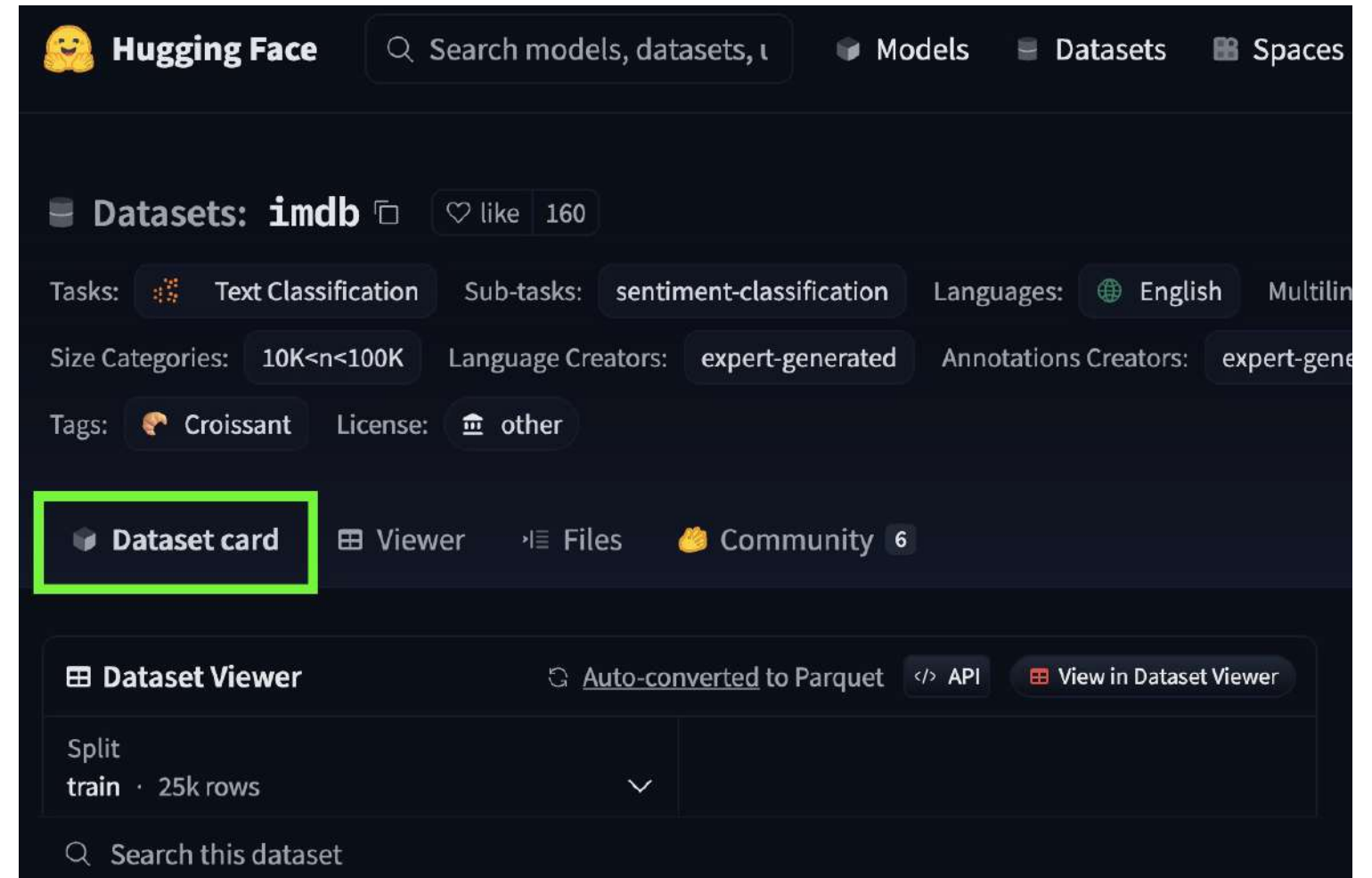
Edit dataset card

<sup>1</sup> <https://huggingface.co/datasets/imdb>



# Dataset cards

- Description
- Dataset structure
- An example
- Field metadata
- Training and testing splits



<sup>1</sup> <https://huggingface.co/datasets/imdb>

# Dataset cards

Datasets: **imdb**

like

160

Dataset card

**Viewer**

Files

Community 6

Split


train · 25k rows

▼

Search this dataset

**text**


string · *lengths*



5213.7k

**label**


class label



2 classes

|   |       |
|---|-------|
| I rented I AM CURIOUS-YELLOW from my video store because of all the controversy that surrounded it when it was first released in 1967. I also heard that at first it was seized by U.S. customs if it ever tried to enter this country,...    | 0 neg |
| "I Am Curious: Yellow" is a risible and pretentious steaming pile. It doesn't matter what one's political views are because this film can hardly be taken seriously on any level. As for the claim that frontal male nudity is an...          | 0 neg |
| If only to avoid making this type of film in the future. This film is interesting as an experiment but tells no cogent story.<br /><br />One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it... | 0 neg |
| This film was probably inspired by Godard's Masculin, féminin and I urge you to see that film instead.<br /><br />The film has two strong elements and those are, (1) the realistic acting (2) the impressive, undeservedly good, photo...    | 0 neg |

<sup>1</sup> <https://huggingface.co/datasets/imdb>

 datacamp

WORKING WITH HUGGING FACE

# Dataset cards

The screenshot shows the Hugging Face website interface for the 'imdb' dataset. The top navigation bar includes the Hugging Face logo, a search bar, and links to Models, Datasets, Spaces, Posts, Docs, and Pricing. The dataset card for 'imdb' is displayed, showing its tasks (Text Classification), sub-tasks (sentiment-classification), languages (English), and multilinguality (monolingual). It also lists size categories (10K<n<100K), language creators (expert-generated), annotations creators (expert-generated), and source datasets (original). The 'Viewer' tab is highlighted with a green box. Below the tabs, the 'Dataset Viewer' section is visible, with the 'View in Dataset Viewer' button highlighted by a green box. The right sidebar shows the download count for the last month as 273,080, along with buttons for 'Use in Datasets library' and 'Edit dataset card'.

**Hugging Face** Search models, datasets, ... Models Datasets Spaces Posts Docs Pricing

**Datasets: imdb** like 160

Tasks: Text Classification Sub-tasks: sentiment-classification Languages: English Multilinguality: monolingual

Size Categories: 10K<n<100K Language Creators: expert-generated Annotations Creators: expert-generated Source Datasets: original

Tags: Croissant License: other

Dataset card **Viewer** Files Community 6

**Dataset Viewer** Auto-converted to Parquet API **View in Dataset Viewer**

Split train · 25k rows

Search this dataset

Downloads last month 273,080

Use in Datasets library

Edit dataset card

<sup>1</sup> <https://huggingface.co/datasets/imdb>

# datasets package

```
pip install datasets
```

- Access
- Download
- Mutate
- Use
- Share

<sup>1</sup> <https://huggingface.co/docs/datasets/index>

# Inspecting a dataset

```
from datasets import load_dataset_builder  
  
data_builder = load_dataset_builder("imdb")
```

```
print(data_builder.info.description)
```

```
Large Movie Review Dataset. This is a dataset for sentiment classification...
```

```
print(data_builder.info.features)
```

```
{'text': Value(dtype='string', id=None), 'label': Value(dtype='string', id=None)}
```

<sup>1</sup> [https://huggingface.co/docs/datasets/load\\_hub](https://huggingface.co/docs/datasets/load_hub)

# Downloading a dataset

```
from datasets import load_dataset  
  
data = load_dataset("imdb")
```

## Split parameter

```
data = load_dataset("imdb", split="train")
```

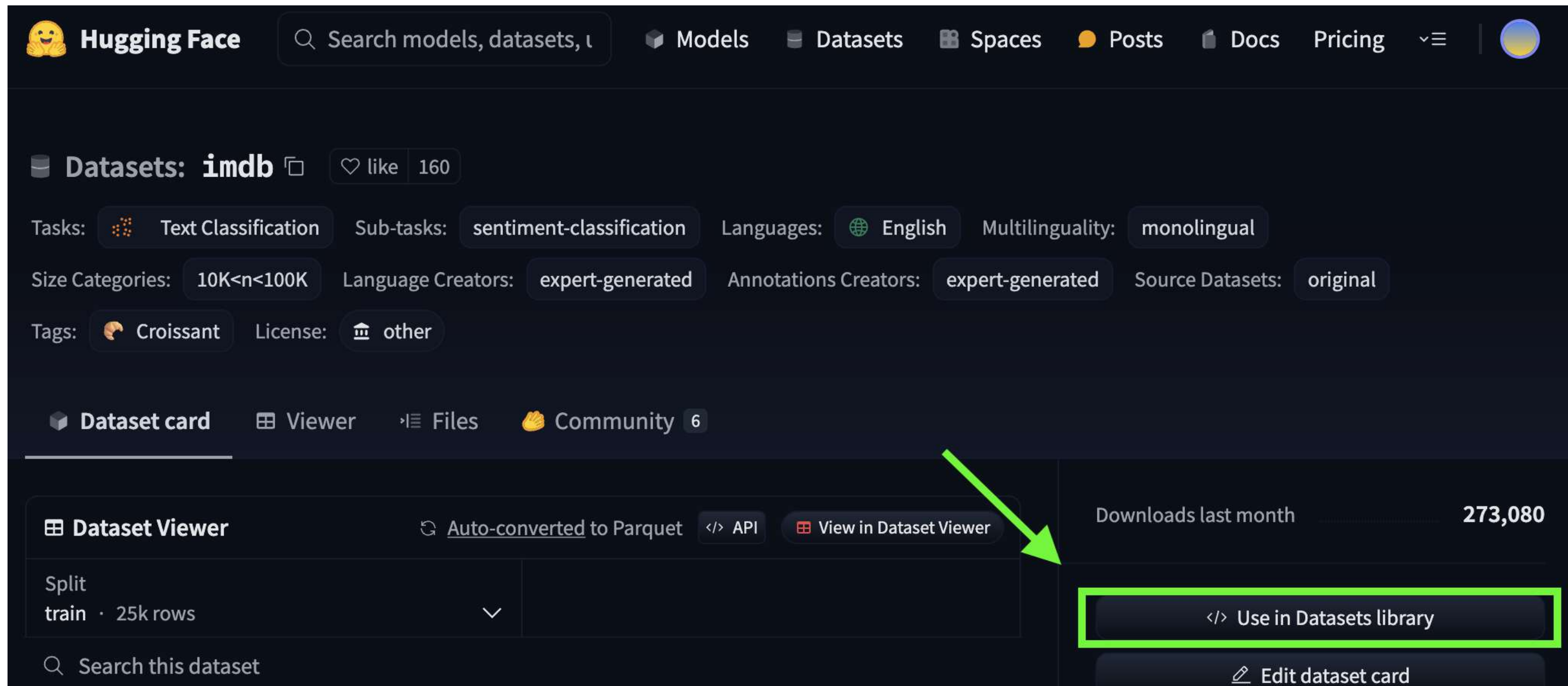
## Configuration parameter

```
data = load_dataset("wikipedia", "20231101.en")
```

<sup>1</sup> <https://huggingface.co/docs/datasets/v2.15.0/loading>



# Use in datasets



The screenshot shows the Hugging Face interface for the 'imdb' dataset. A green arrow points from the 'View in Dataset Viewer' button to the 'Use in Datasets library' button, which is highlighted with a green border. The interface includes a top navigation bar with links to Models, Datasets, Spaces, Posts, Docs, and Pricing. The dataset card for 'imdb' shows it is a Text Classification dataset with 160 likes. It lists various metadata including tasks, sub-tasks, languages, size categories, and tags. The 'Dataset Viewer' section shows the dataset is auto-converted to Parquet and has 25k rows in the train split. The right sidebar displays the download count for the last month as 273,080.

**Hugging Face** Search models, datasets, ... Models Datasets Spaces Posts Docs Pricing

**Datasets: imdb** like 160

Tasks: Text Classification Sub-tasks: sentiment-classification Languages: English Multilinguality: monolingual

Size Categories: 10K<n<100K Language Creators: expert-generated Annotations Creators: expert-generated Source Datasets: original

Tags: Croissant License: other

Dataset card Viewer Files Community 6

**Dataset Viewer** Auto-converted to Parquet API View in Dataset Viewer

Split train · 25k rows

Search this dataset

Downloads last month 273,080

Use in Datasets library

Edit dataset card

# Use in datasets

The screenshot shows the Hugging Face website interface. At the top, there's a navigation bar with the Hugging Face logo, a search bar, and links to Models, Datasets, Spaces, Posts, Docs, and Pricing. Below this, a modal window is open with the title "How to load this dataset with the Datasets library". The modal contains the following code:

```
</> How to load this dataset with the Datasets library
```

```
from datasets import load_dataset

dataset = load_dataset("imdb")
```

There is a "Copy" button next to the code. Below the modal, the dataset card for "imdb" is visible. It includes tabs for "Dataset card", "Viewer", "Files", and "Community". The "Dataset card" tab is selected. The card shows the "Dataset Viewer" section with a split of "train" (25k rows). There are also links for "Auto-converted to Parquet", "API", and "View in Dataset Viewer". The "Downloads last month" are listed as 273,080. At the bottom of the card, there are buttons for "Use in Datasets library" and "Edit dataset card".



# Apache Arrow dataset formats

| Row-based |  |                 | Column-based |  |                 |
|-----------|--|-----------------|--------------|--|-----------------|
| Row 1     |  | 1331246660      | session_id   |  | 1331246660      |
|           |  | 3/8/2012 2:44PM |              |  | 1331246351      |
|           |  | 99.155.155.225  |              |  | 1331244570      |
| Row 2     |  | 1331246351      |              |  | 1331261196      |
|           |  | 3/8/2012 2:38PM | timestamp    |  | 3/8/2012 2:44PM |
|           |  | 65.87.165.114   |              |  | 3/8/2012 2:38PM |
| Row 3     |  | 1331244570      |              |  | 3/8/2012 2:09PM |
|           |  | 3/8/2012 2:09PM |              |  | 3/8/2012 6:46PM |
|           |  | 71.10.106.181   |              |  | 99.155.155.225  |
| Row 4     |  | 1331261196      | source_ip    |  | 65.87.165.114   |
|           |  | 3/8/2012 6:46PM |              |  | 71.10.106.181   |
|           |  | 76.102.156.138  |              |  | 76.102.156.138  |

<sup>1</sup> <https://arrow.apache.org/overview/>

# Mutating a dataset

```
imdb = load_dataset("imdb", split="train")  
  
# Filter imdb  
filtered = imdb.filter(lambda row: row['label']==0)
```

```
{'text': 'I rented I AM CURIOUS-YELLOW...'}
```

<sup>1</sup> <https://huggingface.co/docs/datasets/process#select-and-filter>

# Mutating a dataset

```
# Slicing  
sliced = filtered.select(range(2))
```

```
print(sliced)
```

```
Dataset({features: ['id', 'url', 'title', 'text'], num_rows: 2})
```

```
print(sliced[0]['text'])
```

<sup>1</sup> <https://huggingface.co/docs/datasets/process#select-and-filter>

# Benefits of datasets

- Accessible and shareable
- Relevant to common ML tasks
- Efficient processing on large data
- Faster querying
- Convenient complimentary `datasets` package

# Let's practice!

WORKING WITH HUGGING FACE