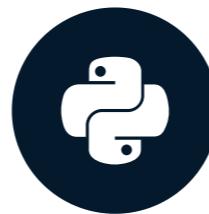


# Introduction to Hugging Face

## 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

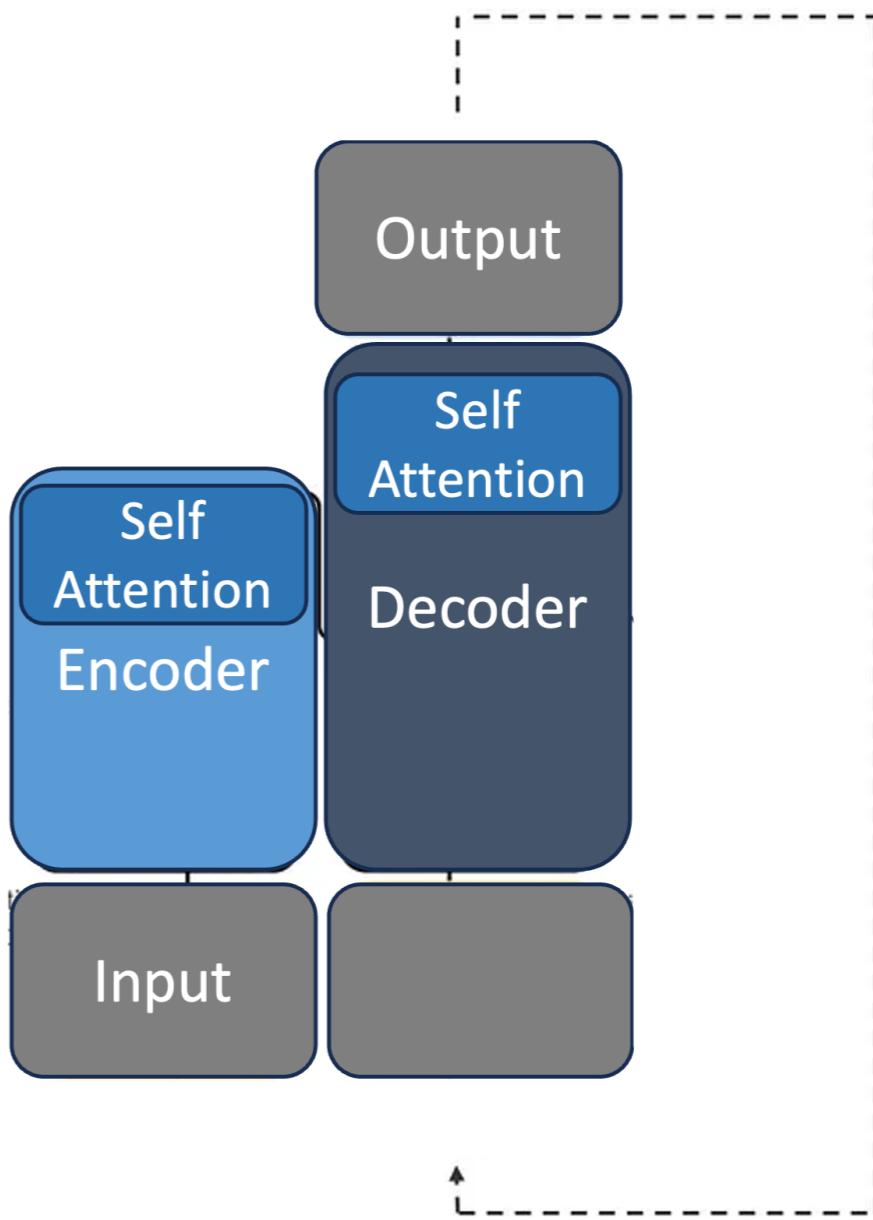
- LLMs
  - Understand and generate human-like text
  - Massive amounts of data
  - Learn patterns in sequences



<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



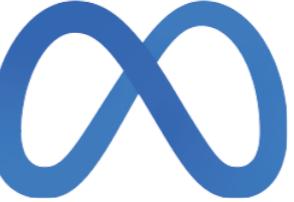
<sup>1</sup> <https://towardsdatascience.com/transformers-89034557de14>

# Large Language Models

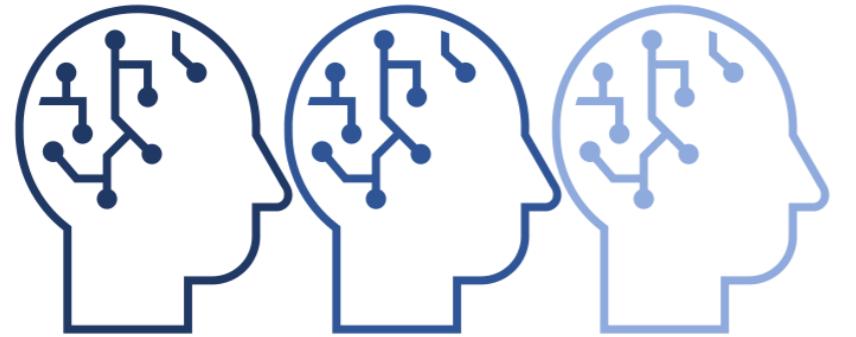
- LLMs
- Understand and generate human-like text
- Massive amounts of data
- Learn patterns in sequences
- Transformer architecture
- Popular options are GPT and Llama



**ChatGPT**

**LLaMA**  
by  Meta

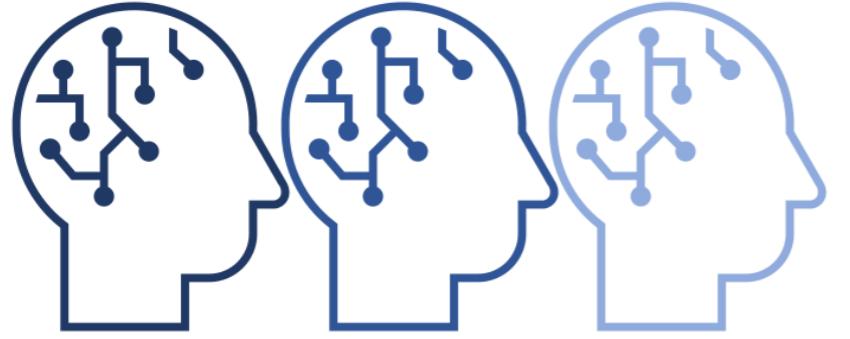
# Benefits of Hugging Face



## Access to Models

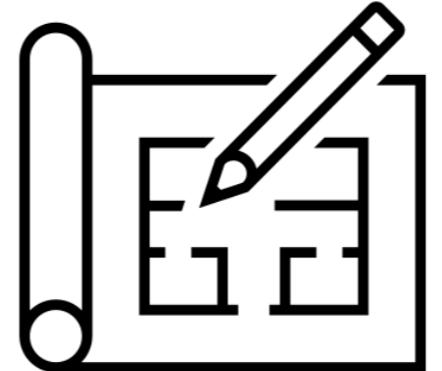
- Faster experimentation

# Benefits of Hugging Face



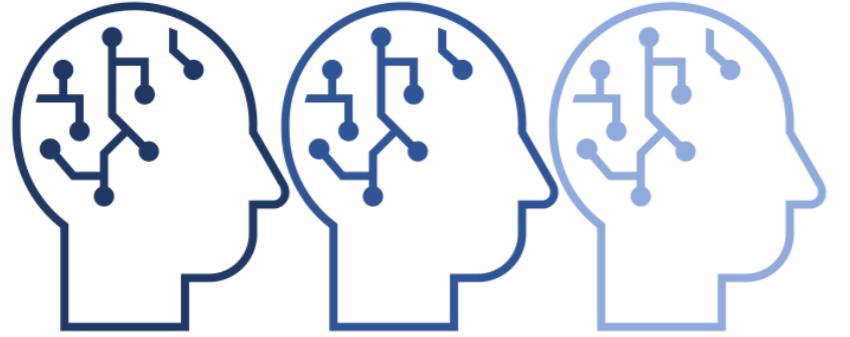
## Access to Models

- Faster experimentation
- Supports every step of the process



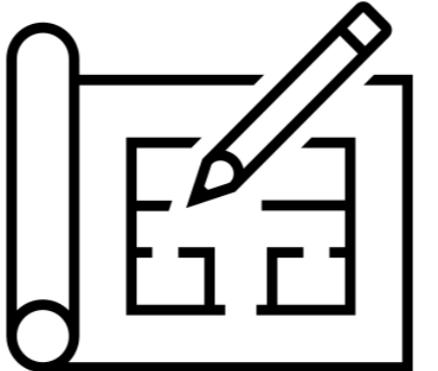
## Frameworks

# Benefits of Hugging Face

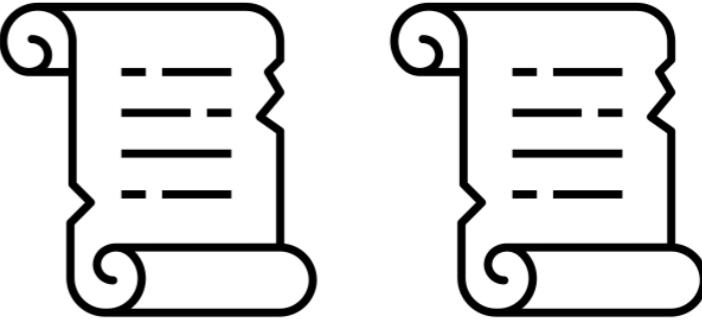


Access to Models

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



Frameworks



Documentation

# 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<sup>1</sup>

```
pip install torch torchvision torchaudio
```

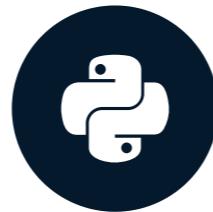
<sup>1</sup> <https://pytorch.org/>

# **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

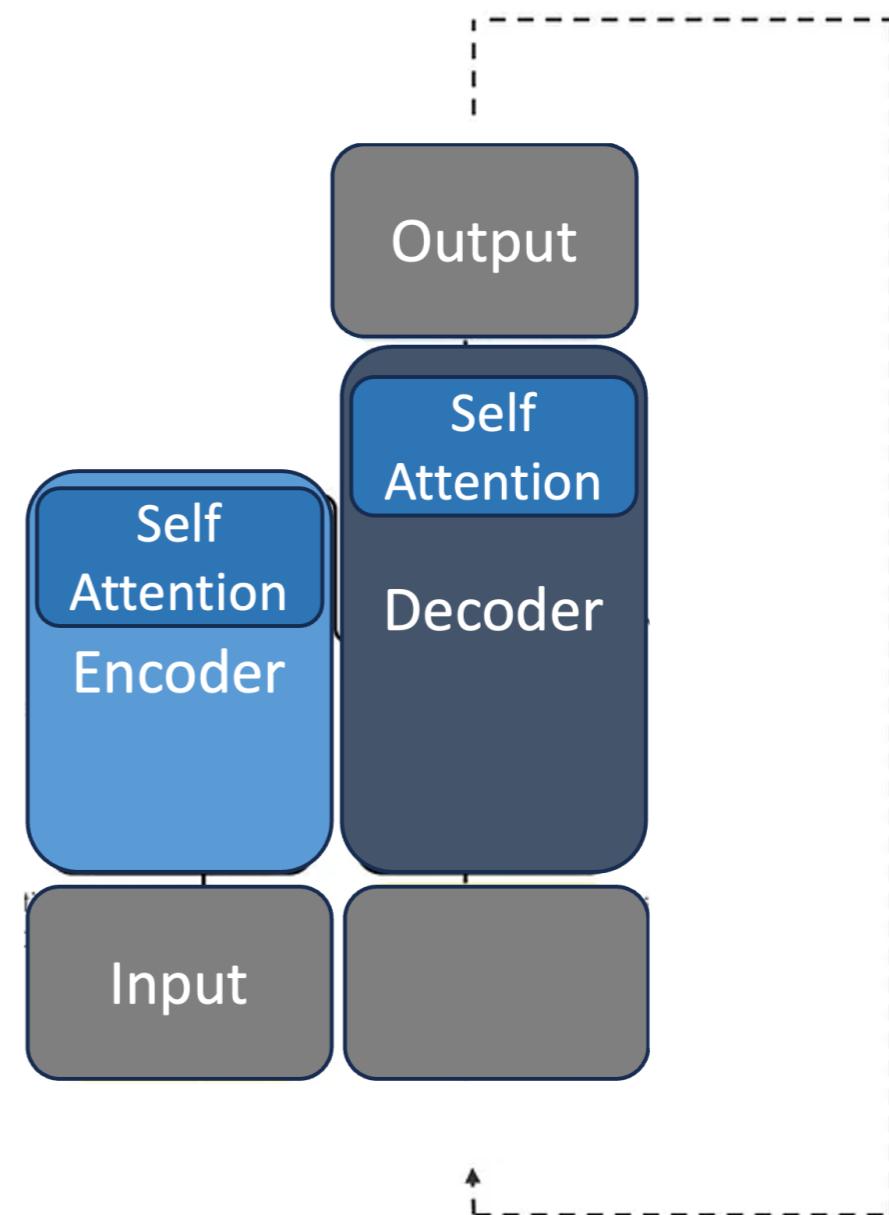


The screenshot shows the GitHub repository page for 'Transformers'. At the top, there's a large yellow emoji of a smiling face with hands clasped. To its right, the word 'Transformers' is written in a large, white, sans-serif font. Below this, there's a horizontal bar containing several status indicators: 'build passing' (green), 'license Apache-2.0' (grey), 'website online' (green), 'release v4.37.2' (blue), 'Contributor Covenant v2.0 adopted' (pink), and 'DOI 10.5281/zenodo.7391177' (blue). Further down, there's a row of language links: English | 简体中文 | 繁體中文 | 한국어 | Español | 日本語 | हिन्दी | Русский | Português | தமிழ் | Français | Deutsch |. Below these, the text 'State-of-the-art Machine Learning for JAX, PyTorch and TensorFlow' is displayed. A large yellow callout box in the center contains a smaller yellow emoji of a graduate wearing a cap, followed by the text 'Part of the Hugging Face course!'. At the bottom of the page, there's a note with a small smiley face emoji: '(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

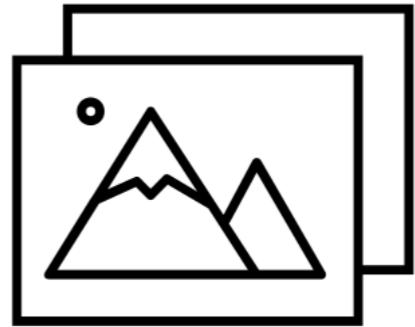
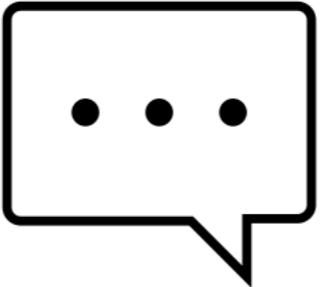
- Neural network models
- Learn context and understanding
- Core components:
  - Encoder
  - Decoder
  - Self-attention mechanism
- Transform input to numerical representations
- Helps model understand context of the input



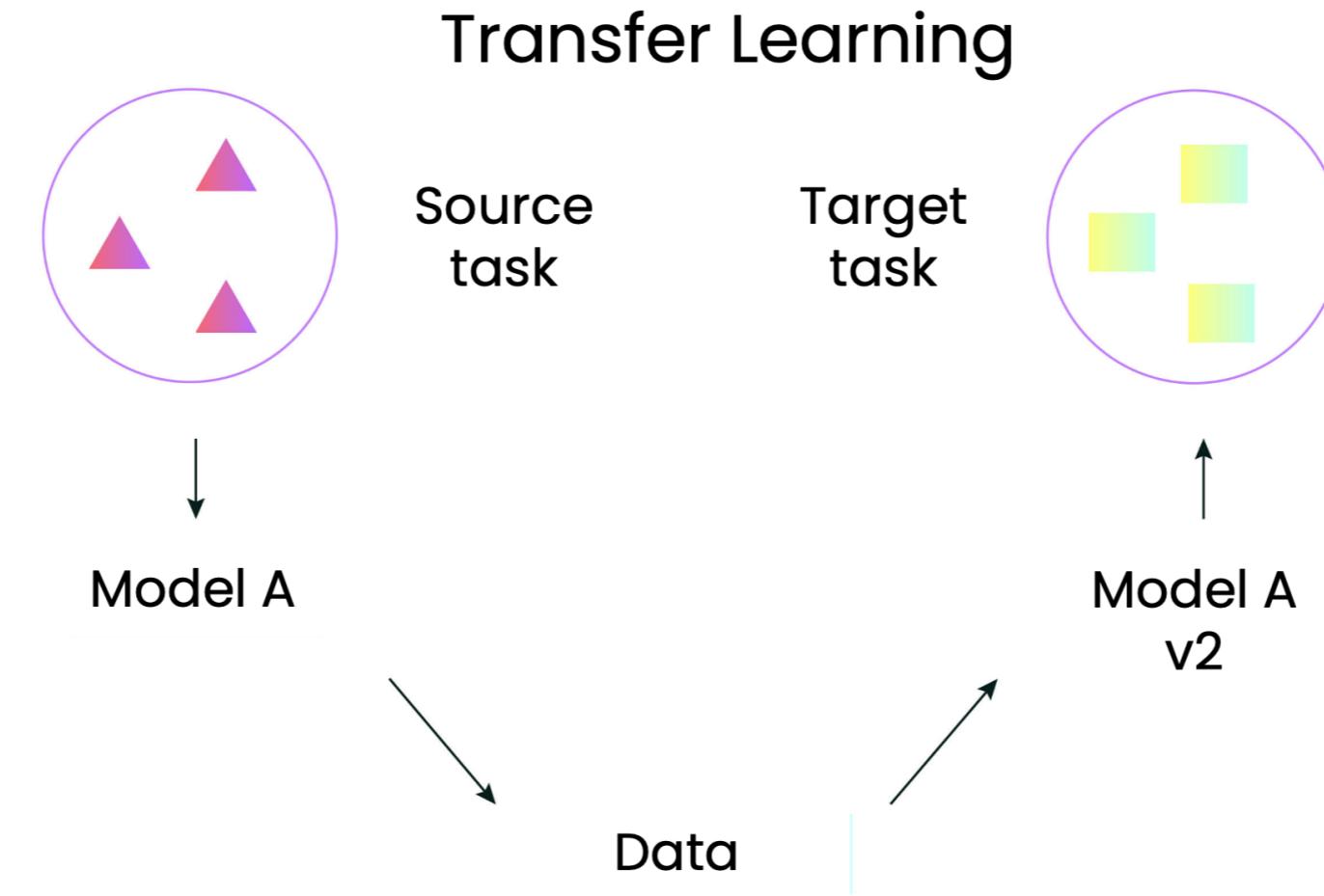
<sup>1</sup> <https://www.turing.com/kb/brief-introduction-to-transformers-and-their-power>

# 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

The screenshot shows the Hugging Face Hub homepage. On the left is a dark sidebar with a yellow smiley face icon, a '+ New' button, and links for Profile, Inbox (0), Settings, Get Pro, Organizations, and Create New. The main area has a search bar, a navigation menu with Models, Datasets, Spaces, Posts, Docs, and Pricing, and a user profile icon. It features sections for 'Following' (with tabs for All, Models, Datasets, Spaces, Papers, Collections, Community, Posts, Upvotes, Likes) and 'Trending' (last 7 days, with tabs for All, Models, Datasets, Spaces). The 'Trending' section lists three models: 'metavoiceio/metavoice-1B-v0.1' (Text-to-Speech, updated 3 days ago, 928 stars, 428 forks), 'briaai/RMBG-1.4' (Image-to-Image, updated about 14 days ago, 68 stars, 383 forks), and 'openbmb/MiniCPM-2B-sft-fp32' (Text Generation, updated 10 days ago, 12.1k stars, 225 forks).

+ New

Profile

Inbox (0)

Settings

Get Pro

Organizations

Create New

Hugging Face

Search models, datasets, users...

Models Datasets Spaces Posts Docs Pricing

Following 0

All Models Datasets Spaces Papers Collections Community Posts

Upvotes Likes

Trending last 7 days

All Models Datasets Spaces

metavoiceio/metavoice-1B-v0.1

Text-to-Speech • Updated 3 days ago • ↓ 928 • ❤ 428

briaai/RMBG-1.4

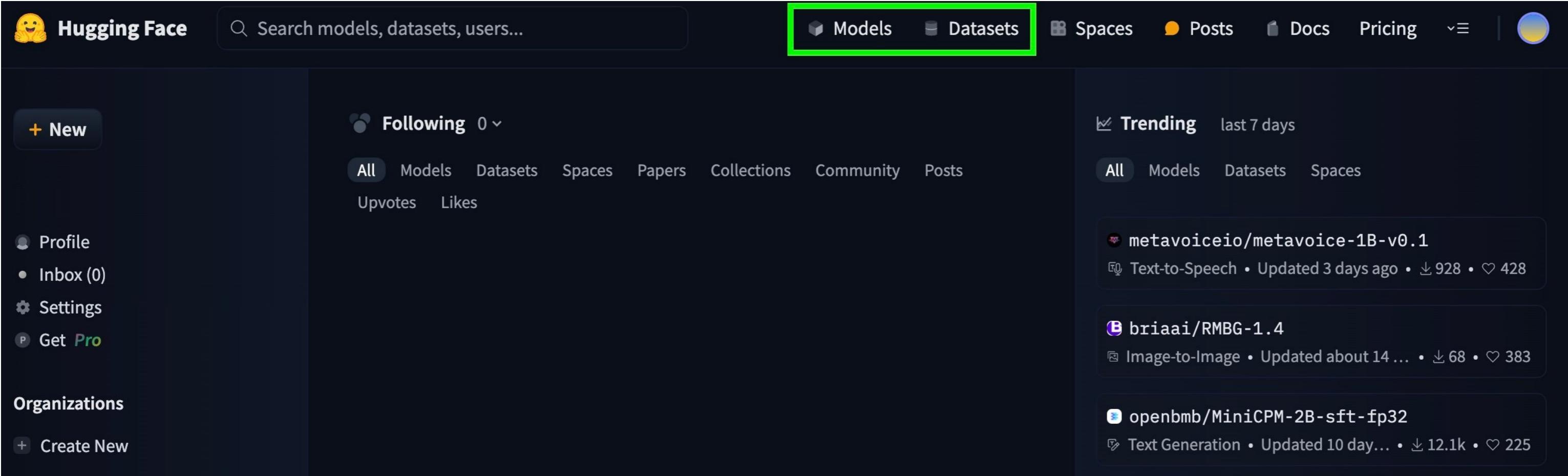
Image-to-Image • Updated about 14 ... • ↓ 68 • ❤ 383

openbmb/MiniCPM-2B-sft-fp32

Text Generation • Updated 10 day... • ↓ 12.1k • ❤ 225

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

# Navigating the Hub



The screenshot shows the Hugging Face Hub homepage. At the top, there is a navigation bar with the Hugging Face logo, a search bar, and several tabs: Models (highlighted with a green box), Datasets, Spaces, Posts, Docs, and Pricing. Below the navigation bar, there are two main sections: 'Following' and 'Trending'. The 'Following' section shows a list of items like 'Profile', 'Inbox (0)', 'Settings', 'Get Pro', and 'Organizations'. The 'Trending' section shows a list of trending models: 'metavoiceio/metavoice-1B-v0.1' (Text-to-Speech, updated 3 days ago, 928 stars, 428 forks), 'briaai/RMBG-1.4' (Image-to-Image, updated about 14 days ago, 68 stars, 383 forks), and 'openbmb/MiniCPM-2B-sft-fp32' (Text Generation, updated 10 days ago, 12.1k stars, 225 forks).

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

# Searching for models

The screenshot shows the Hugging Face website interface. At the top, there is a navigation bar with the Hugging Face logo, a search bar, and several menu items: 'Models' (which is highlighted with a green box), 'Datasets', 'Spaces', 'Posts', 'Docs', and 'Pricing'. Below the navigation bar, there is a secondary navigation bar with links for 'Tasks', 'Libraries', 'Datasets', 'Languages', 'Licenses', and 'Other'. A search bar labeled 'Filter Tasks by name' is also present. On the left side, there is a sidebar titled 'Multimodal' containing various task categories: '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'. On the right side, the main content area displays a list of 505,454 models. The first four models listed are:

- 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. At the top, there is a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets, Spaces, Posts, Docs, Pricing, and a user profile icon. On the left, a sidebar titled "Tasks" is highlighted with a green box. It contains categories like Libraries, Datasets, Languages, Licenses, and Other, along with a "Filter Tasks by name" search bar. Below this, under "Multimodal", are several buttons for 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. On the right, the main content area displays a list of 505,454 models. The first four results are shown:

- 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. At the top, there's a navigation bar with icons for Models, Datasets, Spaces, Posts, Docs, Pricing, and a user profile. Below the navigation is a search bar and a main menu with options like Tasks, Libraries, Datasets, Languages (which is highlighted), and Licenses. A secondary navigation bar below the main menu includes a 'Filter Languages by name' search bar and a grid of language buttons: 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. To the right, a large section displays 'Models 505,454'. It features a 'Filter by name' button and sorting options ('Full-text search', 'Sort: Trending'). Below this, four model cards are listed:

- 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. At the top, there's a navigation bar with icons for Models, Datasets, Spaces, Posts, Docs, Pricing, and a user profile. Below the navigation is a search bar labeled "Search models, datasets, ...". On the left, there's a sidebar with tabs for Tasks, Libraries (which is highlighted with a green box), Datasets, Languages, and Licenses. Under the Libraries tab, there's a "Filter Libraries by name" input field and a grid of library names with their respective logos: 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.

**Models 505,454**

Filter by name  Full-text search

**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

Screenshot of the Hugging Face Model card for `openai/whisper-large-v3`.

The page includes the following elements:

- Hugging Face logo** and **Search bar** at the top.
- Model card title**: `openai/whisper-large-v3` with **1.63k likes**.
- Tags and categories**:
  - 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
- Navigation tabs**: Model card, Files, Community (80), Train, Deploy, Use in Transformers.
- Summary section**:
  - Whisper**: 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 (represented by a purple line chart).
  - 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 datasets homepage. At the top, there is a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets (which is highlighted with a green box), Spaces, Posts, Docs, Pricing, and a user profile icon. Below the navigation bar, there are filters for Tasks (Tasks, Sizes, Sub-tasks, Languages, Licenses, Other), a search bar for filtering tasks by name, and sections for Multimodal and Computer Vision tasks. The main content area displays a list of datasets with their names, last updated time, viewer count, and star count. The first dataset listed is Locutusque/UltraTextbooks.

Dataset	Last Updated	Viewers	Stars
Locutusque/UltraTextbooks	11 days ago	276	136
teknum/OpenHermes-2.5	8 days ago	6.59k	368
allenai/dolma	11 days ago	665	584
fka/awesome-chatgpt-prompts	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 for searching datasets. The top navigation bar includes the Hugging Face logo, a search bar, and links for Models, Datasets, Spaces, Posts, Docs, Pricing, and user profile.

The main search interface features a navigation bar with tabs: Tasks (highlighted with a green border), Sizes, Sub-tasks, Languages, Licenses, and Other. Below this is a search bar labeled "Filter Tasks by name".

On the left, there are two sections: "Multimodal" and "Computer Vision". The "Multimodal" section lists categories: Feature Extraction, Text-to-Image, Image-to-Text, Image-to-Video, Text-to-Video, Visual Question Answering, Graph Machine Learning, Text-to-3D, and Image-to-3D. The "Computer Vision" section lists Depth Estimation and Image Classification.

The right side displays a list of datasets:

- Locutusque/UltraTextbooks**  
Viewer • Updated 11 days ago • ↓ 276 • ❤ 136
- teknum/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 dataset card for 'imdb'. At the top, there's a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets, Spaces, Posts, Docs, Pricing, and a user profile icon. Below the navigation bar, the dataset title 'Datasets: imdb' is displayed with a like count of 160. The dataset details section includes fields for 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), and License (other). Below this, there are tabs for Dataset card, Viewer, Files, and Community (with 6 items). The Dataset Viewer tab is active, showing a preview of the dataset with a split of 'train' (25k rows) and a 'Search this dataset' input field. To the right of the viewer, there are statistics: 'Downloads last month' (273,080) and buttons for 'Use in Datasets library' and 'Edit dataset card'.

Datasets: **imdb** 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

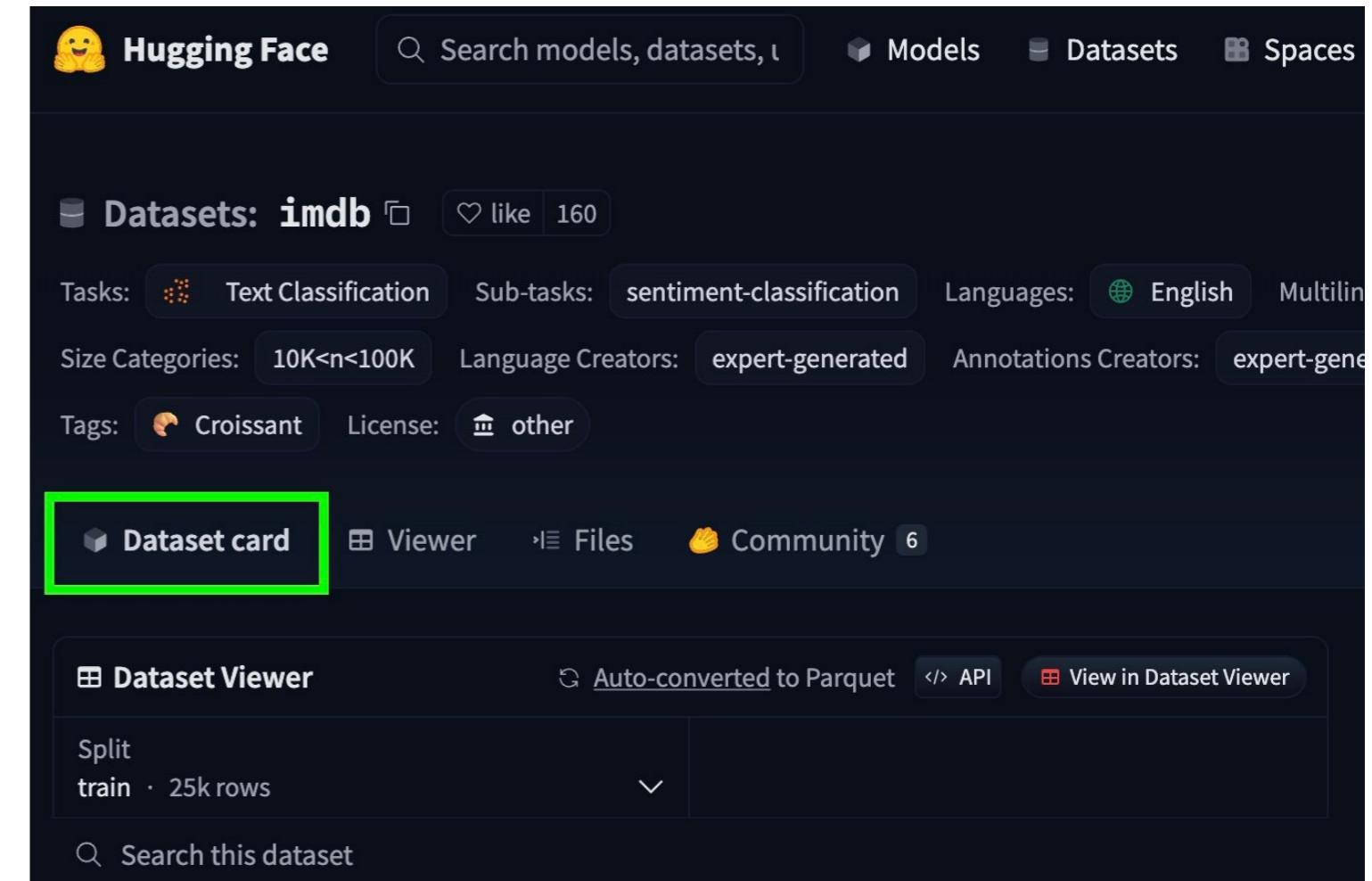
Downloads last month **273,080**

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	label
string · lengths 52 13.7k	class label 2 classes 0 neg
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.  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.  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>

# Dataset cards

The screenshot shows the Hugging Face dataset card for 'imdb'. At the top, there's a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, Pricing, and a user profile icon. Below the navigation bar, the dataset title 'Datasets: imdb' is displayed, along with a 'like' button showing 160 likes. The dataset details include:

- 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

Below the details, there are tabs for Dataset card (selected), Viewer (highlighted with a green box), Files, and Community. The 'Viewer' tab shows the dataset structure with a 'Dataset Viewer' section containing a 'Split' table (train · 25k rows) and a 'Search this dataset' input field. Other sections include 'Auto-converted to Parquet', 'API', 'View in Dataset Viewer' (highlighted with a green box), 'Downloads last month' (273,080), 'Use in Datasets library', and '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 Datasets interface for the 'imdb' dataset. At the top, there's a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, Pricing, and a user profile icon. Below the navigation, the dataset card for 'imdb' is displayed, showing its details: 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, and License: other. Below the card, there are tabs for Dataset card, Viewer, Files, and Community. The 'Dataset Viewer' tab is active, showing a preview of the dataset with a split of 'train · 25k rows'. There are buttons for Auto-converted to Parquet, API, View in Dataset Viewer, and Use in Datasets library. A large green arrow points from the 'View in Dataset Viewer' button to the 'Use in Datasets library' button. To the right, it says 'Downloads last month: 273,080'. At the bottom, there's a search bar for 'Search this dataset' and an 'Edit dataset card' button.

# Use in datasets

The screenshot shows a dark-themed interface for the Hugging Face Datasets library. At the top, there's a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, and Pricing. Below the navigation, a modal window is open with the title '</> How to load this dataset with the • Datasets ⓘ library'. It contains two code snippets:

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

There's a 'Copy' button next to the second snippet. Below the modal, the main page content includes:

- A 'Dataset card' tab is selected.
- Other tabs include 'Viewer', 'Files', and 'Community' (with 6 items).
- A 'Dataset Viewer' section shows the dataset has been auto-converted to Parquet format.
- Statistics: Downloads last month (273,080).
- A 'Split' section indicates a train set with 25k rows.
- Links: 'Use in Datasets library' and 'Edit dataset card'.
- A search bar at the bottom left says 'Search this dataset'.

# 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
	1331246351		1331261196
Row 2	3/8/2012 2:38PM	timestamp	3/8/2012 2:44PM
	65.87.165.114		3/8/2012 2:38PM
	1331244570		3/8/2012 2:09PM
	3/8/2012 2:09PM		3/8/2012 6:46PM
Row 3	71.10.106.181	source_ip	99.155.155.225
	1331261196		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**