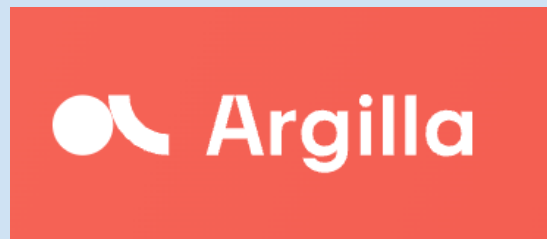


# *Creating Synthetic Dataset for Text Classification using Ollama*



X



**Hugging Face**

X



# Introduction

In the **electrical engineering field**, **sentiment analysis** has gained significant importance as industries seek to harness **customer feedback**, **technical reports**, and **stakeholder insights**. However, collecting sufficient labeled data for **text classification** in this domain can be challenging due to its technical nature and the scarcity of domain-specific datasets, which are often confidential. **Synthetic dataset generation** addresses this gap by enabling the creation of tailored datasets that are **cost-efficient**, **privacy-conscious**, and adaptable to evolving needs. It ensures that sentiment models are robust and well-equipped to handle the unique challenges of **electrical engineering applications**.

Thanks to **Argilla** and its team, a robust solution for generating synthetic datasets is now available. Recently, David created a **synthetic generator app** that simplifies the process of creating synthetic datasets in no time.

## Features of the Synthetic Generator App

The app supports multiple ways to access **LLMs** (Large Language Models), including **Anthropic**, **OpenAI**, **Hugging Face Inference**, **VertexAI**, **Ollama**, and more. In this exercise, we demonstrate how **Ollama** can be integrated with the app to create a custom synthetic dataset at no cost.

## How to Set Up the Synthetic Generator App

For this guide, we will use the **Llama 3.1:8B model** to generate our dataset.

### Prerequisites

1. **Install Ollama** on your local machine.
2. Pull the model to be used from the Ollama hub. For example, to use **Llama 3.1:8B**, run: `ollama run llama3.1`

### Steps to Install Dependencies

1. Clone the repository into a folder of your choice: `git clone https://github.com/argilla-io/synthetic-data-generator.git`
2. Create and activate a Conda environment:  
`conda create -n synthetic-dataset python=3.12`  
`conda activate synthetic-dataset`

3. Install the dependencies: `pip install -e .` Also, install the `python-dotenv` package: `pip install python-dotenv`

## Running the Synthetic Dataset Generator with Ollama

To run the app with the Ollama model, copy and paste the following script into `app.py`:

```
import os
from dotenv import load_dotenv

_ = load_dotenv()

from synthetic_dataset_generator import launch

assert os.getenv("HF_TOKEN")

os.environ["BASE_URL"] = "http://127.0.0.1:11434/v1/"
os.environ["MODEL"] = "llama3.1"
os.environ["MAX_NUM_ROWS"] = "20000"

launch(share=True)
```

Refer to the **README** file of the repository to check which environment variables can be customized to modify the generation process.

## Setting the Hugging Face Token

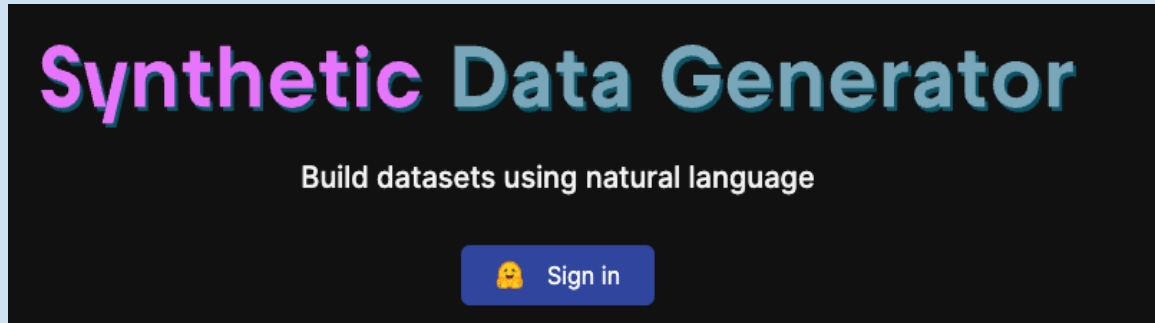
1. Obtain a **Hugging Face token** in write mode from your profile.
2. Set the token as an environment variable in the `.env` file:

```
HF_TOKEN=<your_hugging_face_token>
```

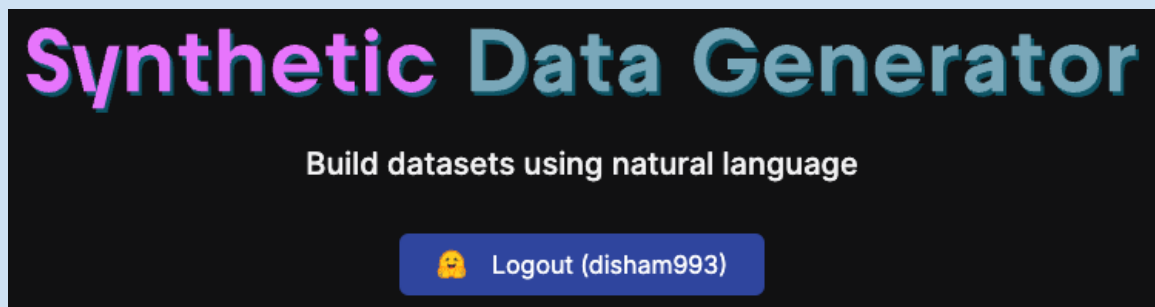
You are now ready to run the app: `python app.py`

## Generating Synthetic Data for Text Classification

1. Once the app is running, copy and paste the link from the terminal into your browser. Click on **Sign In** to log into your Hugging Face profile automatically. Before sign-in



2. After sign-in:



## Workflow for Dataset Generation

- **Describe the Dataset:** Provide a description, and the app will automatically generate a system prompt and labels.

The screenshot shows the 'Text Classification' tab selected. Under '1. Describe the dataset you want', there is a 'Dataset description' box containing the text: 'A dataset containing customer feedback on various electrical devices, classified into sentiment categories such as Positive, Neutral, Negative, and Mixed.' Below this box are 'Clear' and 'Create' buttons. To the right, under 'Examples', there are two example dataset descriptions: 'A dataset covering customer reviews for an e-commerce website.' and 'A dataset covering news articles about various topics.'

The screenshot shows the '2. Configure your dataset' step. On the left, there are three sections: 'System prompt' with the text 'Classify customer feedback on electrical devices into one of the following categories', 'Labels' with a list of 'negative', 'neutral', 'positive', and 'mixed' tags, and 'Clarity' with a dropdown menu set to 'Mixed'. Below these is a 'Difficulty' dropdown also set to 'Mixed'. On the right, a table displays the generated labels and text examples.

label	text
	the display's brightness to be somewhat lacking even at its highest setting.
positive	I've been using this smart plug for my living room lamps for a week now, and it's been amazing so far. The Wi-Fi connectivity is stable and the energy usage monitoring feature is really helpful in reducing our electricity bill.
positive	I was able to set up the new smart plug in about 10 minutes, which is impressive considering I'm not tech-savvy.
negative	I'm extremely disappointed in the new smart plug I bought from your store. The device is constantly disconnecting from Wi-Fi and I've tried resetting it multiple times, but nothing seems to be working.
positive	I was extremely satisfied with my new smart plug. It works flawlessly and the energy monitoring feature is so helpful in keeping track of my usage.
negative	I'm extremely disappointed in the new portable generator I purchased from your company. It's supposed to be lightweight and compact, but it weighs over 50 pounds and takes up way too much space when you're trying to transport it.
negative	This power strip has several USB ports, but the two AC outlets are positioned in such a way that they're hard to access when using the USB ports. Also, the cord storage is very

- **Manual Configuration:** Alternatively, create a system prompt and labels manually.

## Configurations

- **Clarity Options:** Choose from "Clear," "Understandable," "Ambiguous," or "Mixed." Select "Mixed" for balanced sampling.
- **Comprehension Level:** Set the comprehension level to "High School," "College," "Mixed," or "Ph.D." For this exercise, we select "Mixed."

## Sample Generation

- Click **Save** to generate a sample of 10 rows to validate the direction.

**2. Configure your dataset**

**System prompt**

You are a helpful assistant. Your task is to classify customer feedback for various electrical devices (e.g., circuit breakers, transformers, smart meters, inverters, solar panels, and power strips) based on sentiment into one of the following categories: Positive, Neutral, Negative, or Mixed.

**Labels**

Add the labels to classify the text.

Positive x Neutral x Negative x Mixed x

If checked, the text will be classified into multiple labels.

☐ Multi-label

**Clarity**

Set how easily the correct label or labels can be identified.

Mixed

**Difficulty**

Select the comprehension level for the text. Ensure it matches the task context.

Mixed

Clear Save

label	text
positive	The circuit breaker's ability to self-test and reset without requiring manual intervention is a game-changer for industrial settings where downtime needs to be minimized.
negative	The recent transformer installation at our substation exhibited anomalous thermal behavior due to an unforeseen combination of ambient temperature fluctuations and inherent design parameterization.
negative	The smart meter installed at my house exhibits erratic behavior in temperatures below -10°C, occasionally displaying incorrect energy readings.
negative	The implementation of the smart meter's bidirectional communication protocol resulted in unexpected oscillations in the grid frequencies during peak usage periods due to the lag introduced by the meter's proprietary software update.
mixed	The new circuit breaker installation at our office building has been a mixed blessing. On one hand, the smart features have improved energy efficiency and reduced costs by 15%. However, some of the employees are experiencing occasional false tripping incidents due to electromagnetic interference from neighboring devices.
mixed	The circuit breaker's sensitivity adjustment mechanism is ingenious, however, it sometimes interferes with the smart meter's data transmission signals.

- Once satisfied, set the repository name and the number of rows (not exceeding the maximum rows set in the environment variable).
- Click **Push to Hub** to start the dataset generation process.

**3. Generate your dataset**

**Organization**

disham993

**Repo name**

ElectricalDeviceFeedback

**Number of rows**

10000

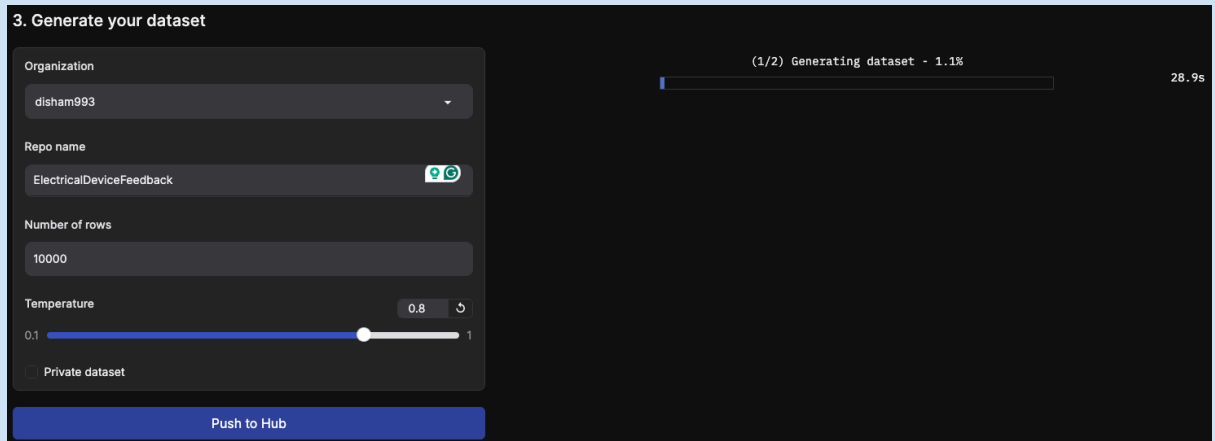
**Temperature**

0.1 0.8 1

☐ Private dataset

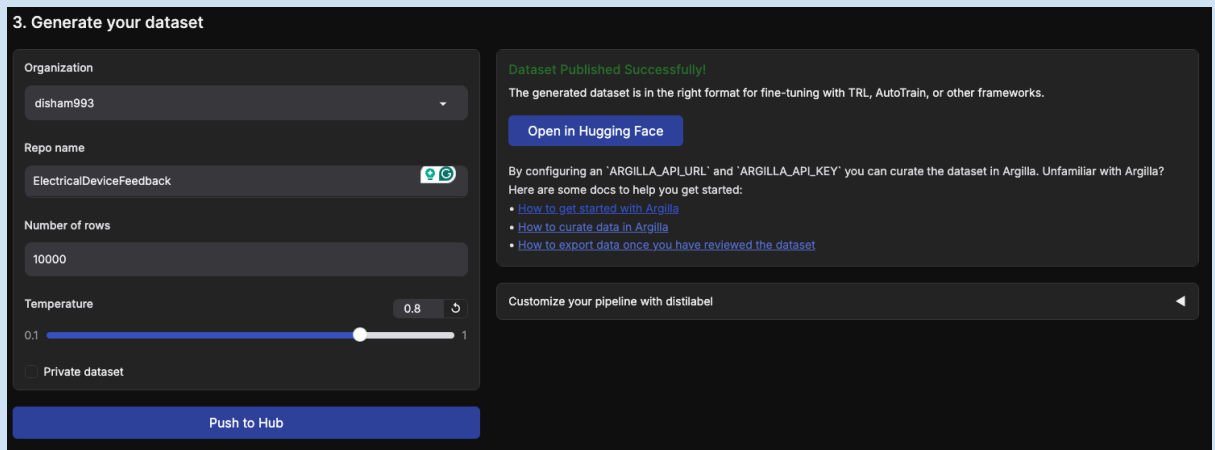
Push to Hub

- Dataset generation started.



The screenshot shows the '3. Generate your dataset' interface. On the left, there are input fields for 'Organization' (disham993), 'Repo name' (ElectricalDeviceFeedback), 'Number of rows' (10000), and 'Temperature' (0.8). A 'Push to Hub' button is at the bottom. On the right, a progress bar indicates '(1/2) Generating dataset - 1.1%' with a timer showing '28.9s'.

- Once the dataset generation is complete, you will see the following message:



The screenshot shows the '3. Generate your dataset' interface after completion. The left sidebar remains the same. The main area displays a green message: 'Dataset Published Successfully!'. Below this, it states 'The generated dataset is in the right format for fine-tuning with TRL, AutoTrain, or other frameworks.' and provides a button 'Open in Hugging Face'. There are also links for 'How to get started with Argilla', 'How to curate data in Argilla', and 'How to export data once you have reviewed the dataset'. A section titled 'Customize your pipeline with distilabel' is also visible.

## Conclusion

The rapid advancements in **LLMs** and supporting technologies have significantly simplified **synthetic dataset generation**. This approach offers a **cost-effective** and efficient way to create datasets tailored to specific needs, ensuring robust and scalable solutions for **text classification** in technical domains such as **electrical engineering**. This guide highlights how tools like the **Synthetic Generator App** and **Ollama** can make the process straightforward and highly adaptable.