Creating Synthetic Dataset for Text Classification using Ollama

Argilla

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

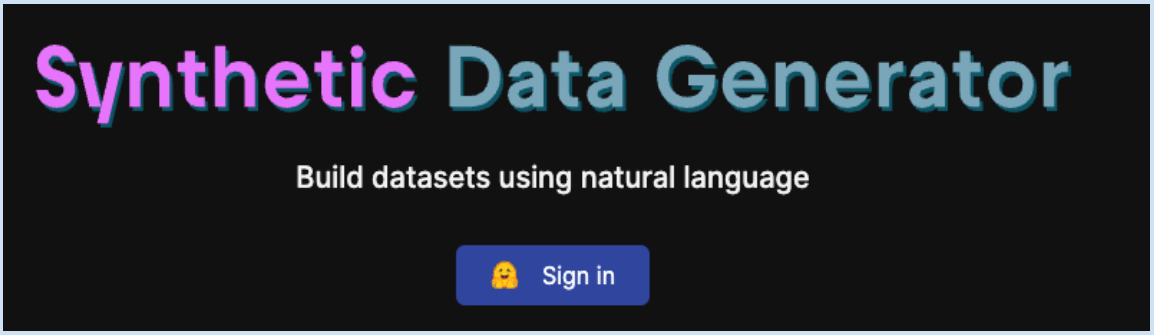
1. Obtain a **Hugging Face token** in write mode from your profile.
2. Set the token as an environment variable in the `.env` file:

   `HF_TOKEN=<your_hugging_face_token>`

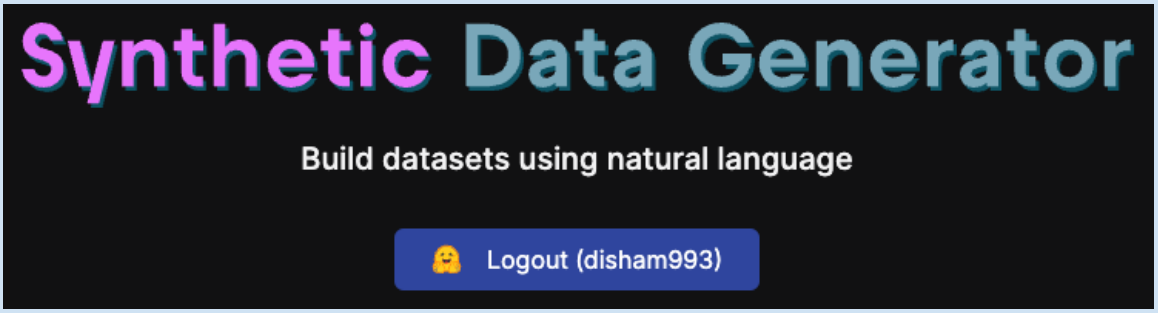
You are now ready to run the app: `python app.py`

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

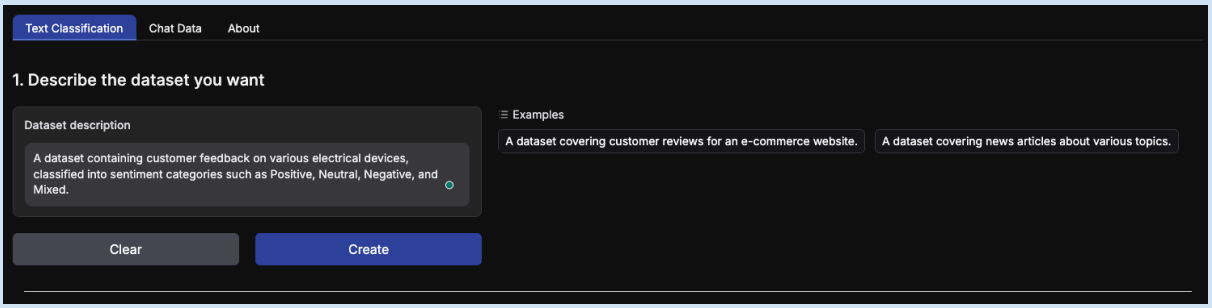


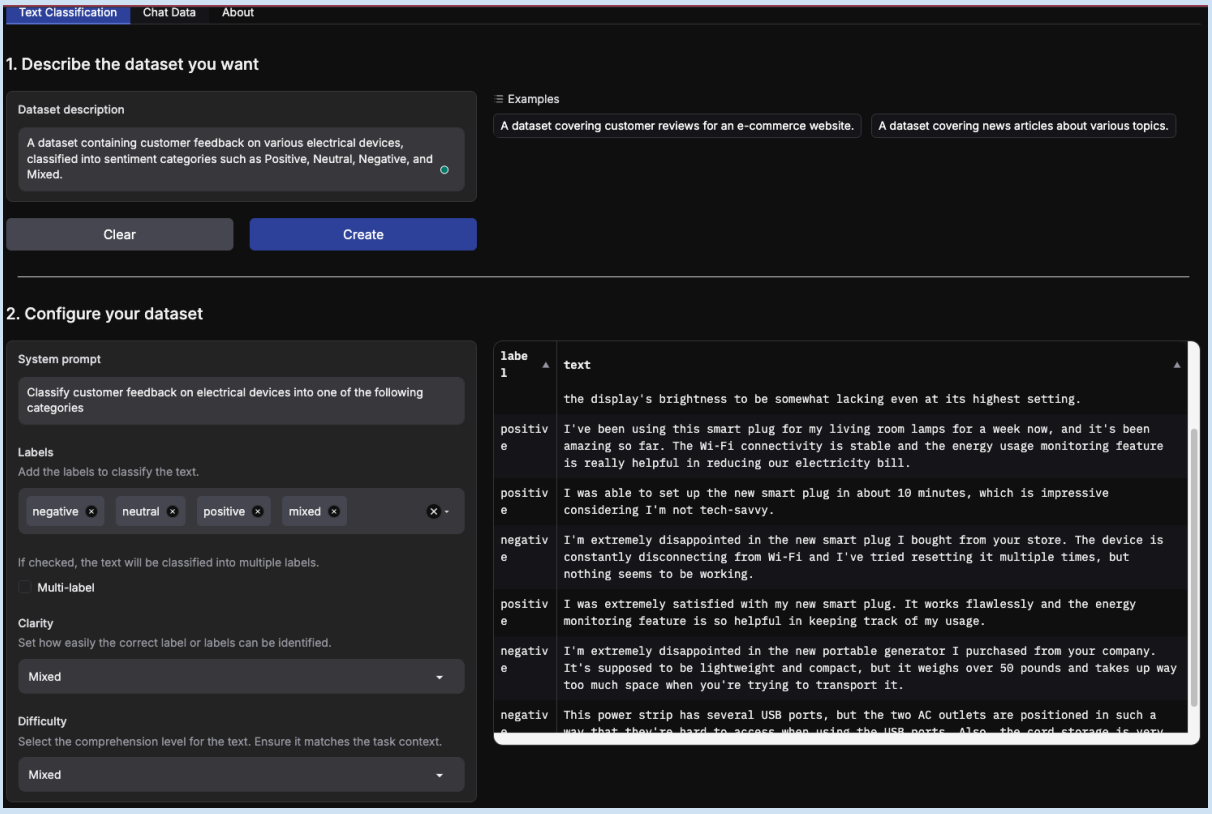
2. After sign-in:



Workflow for Dataset Generation

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





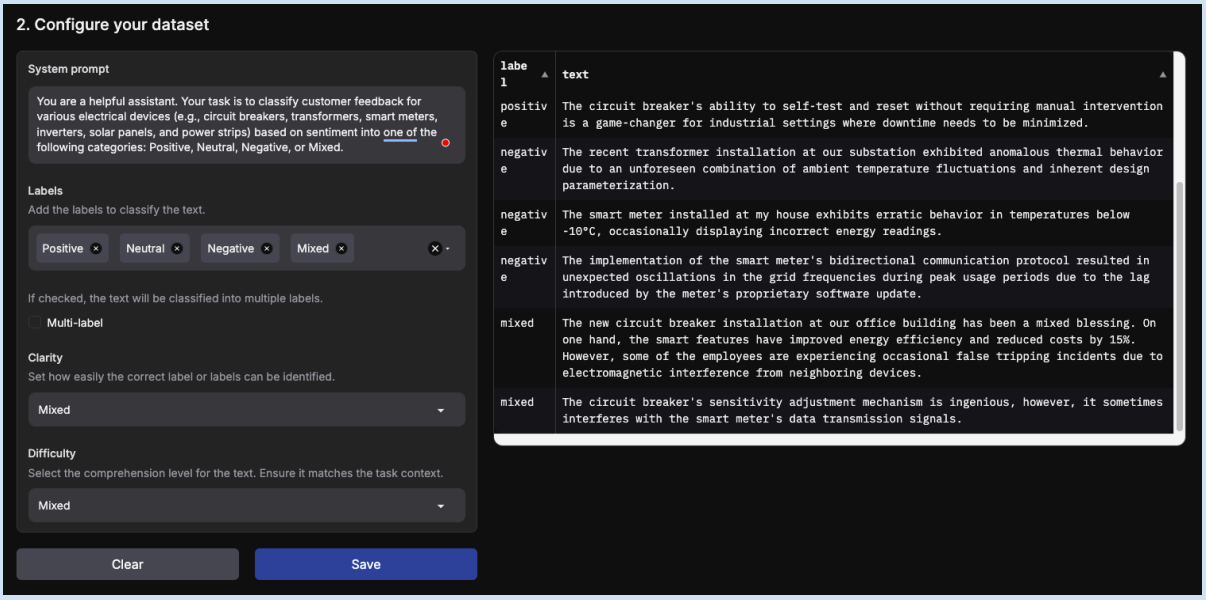
- **Manual Configuration:** Alternatively, create a system prompt and labels manually.

Configurations

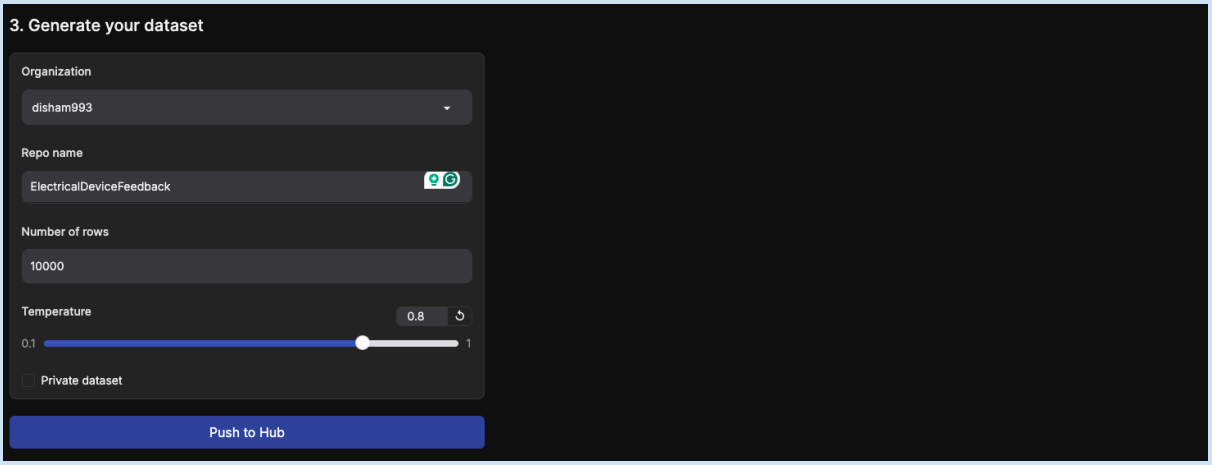
- **Clarity Options:** Choose from "Clear," "Understandable," "Ambiguous," or "Mixed." Select "Mixed" for balanced sampling.
- **Comprehension Level:** Set the comprehension level to "High School," "College," "Mixed," or "Ph.D." For this exercise, we select "Mixed."

Sample Generation

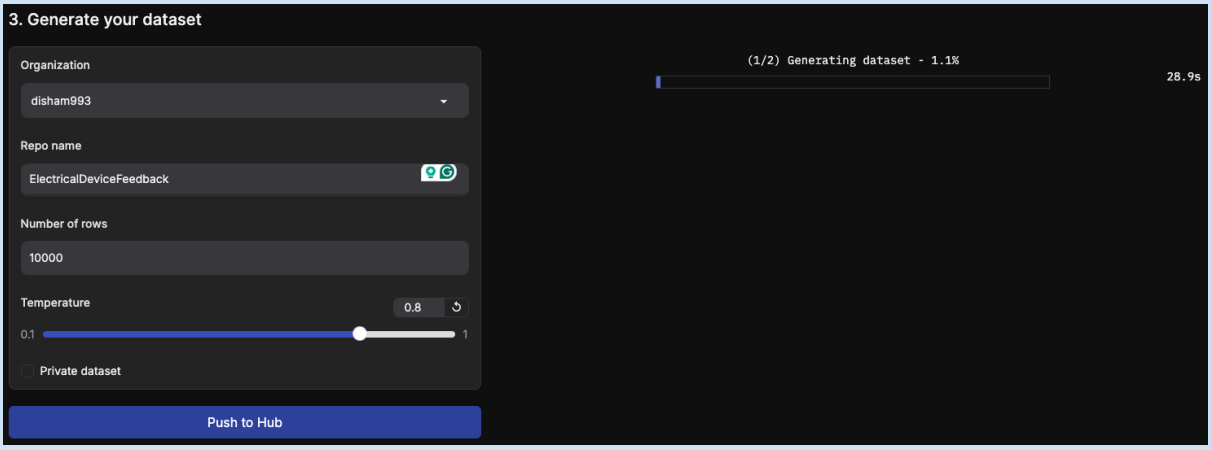
- Click **Save** to generate a sample of 10 rows to validate the direction.



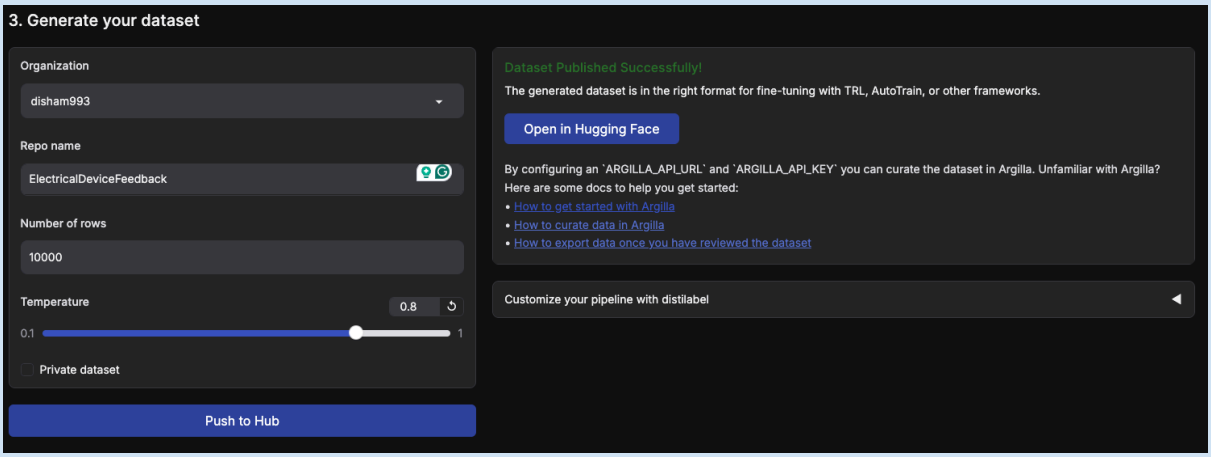
- Once satisfied, set the repository name and the number of rows (not exceeding the maximum rows set in the environment variable).
- Click **Push to Hub** to start the dataset generation process.



- Dataset generation started.



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



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.