README for Person 2: Dataset Processing & Text Classification (Stream 2)

Understood! Here's a comprehensive and detailed step-by-step guide, including terminal commands, for your project. You can use these instructions as a base and ask for further details if needed.

1. Dataset Download & Processing

Download the Ruby Dataset

1. **Access the Dataset:**
    - Open Google Drive link in your web browser.
    - Click "Download" to save the file.
2. **Check File Format:**
    - Determine if the file is in CSV, JSON, or text format.

Transform the Dataset for Text-to-Speech

1. **Install Pandas Library:**

Open your terminal and run:
bash
Copy code
```
pip install pandas
```

    -
2. **Create a Python Script:**

Create a file named `process_data.py` with the following content:
python
Copy code
```
import pandas as pd

# Load the dataset
df = pd.read_csv('path_to_your_downloaded_dataset.csv')  # Replace
with your file path

# Clean the text data
df['text'] = df['text'].str.replace(r'[^\w\s]', '', regex=True)

# Save cleaned text
df['text'].to_csv('cleaned_text.txt', index=False, header=False)
```

○

3. **Run the Python Script:**

Execute the script with:
bash
Copy code
```
python process_data.py
```

○

2. Text Classification

Install Hugging Face Transformers

1. **Install the Transformers Library:**

Run in your terminal:
bash
Copy code
```
pip install transformers
```

○

2. **Create a Classification Script:**

Create a file named `classify_text.py` with:
python
Copy code
```
from transformers import pipeline

# Initialize the classifier
classifier = pipeline('text-classification')

# Load transcribed text
with open('transcribed_text.txt', 'r') as file:
    transcribed_text = file.read()

# Classify the text
results = classifier(transcribed_text)

# Print results
print(results)
```

○

3. **Run the Classification Script:**

Execute the script with:
bash
Copy code
```
python classify_text.py
```

   -

3. Testing & Refinement

Test and Evaluate Accuracy

1. **Prepare Sample Texts:**
   - Create text files with sample data. Place them in the same directory as your scripts.
2. **Run Tests:**

Replace `transcribed_text.txt` with your sample files and rerun the classification script:
bash
Copy code
```
python classify_text.py
```

   -
3. **Evaluate Results:**
   - Review the output for accuracy.

Make Necessary Adjustments

1. **Tune Model Parameters:**
   - Modify parameters in the script if needed. For Hugging Face models, you may adjust the pipeline configuration.
2. **Refine the Process:**
   - Update your scripts based on test results and re-test.