**Hindi-English Translator App**

**1. Project Overview**

This is a **Hindi-English bilingual translation** web application built using **Streamlit**, **Hugging Face Transformers**, and **Hugging Face Datasets**.  
The app provides two translation methods:

* **Neural Machine Translation** using pre-trained models (Helsinki-NLP/opus-mt-hi-en and Helsinki-NLP/opus-mt-en-hi).
* **Basic Lookup Translation** using a dictionary created from the **cfilt/iitb-english-hindi** dataset.

**2. Key Features**

* **Translate from Hindi to English** or **English to Hindi**.
* **Choose Translation Method**:
  + **Neural Model (Higher Accuracy)**: Uses advanced deep learning models.
  + **Basic Lookup (Limited Accuracy)**: Simple word-by-word or sentence matching from a small dataset.
* **Dictionary-based Translation** with sample viewing.
* **Custom Styled Interface** using HTML and CSS within Streamlit.

**3. Technologies Used**

| **Technology** | **Purpose** |
| --- | --- |
| **Streamlit** | Building the web interface |
| **Hugging Face Transformers** | Loading translation models |
| **Hugging Face Datasets** | Loading English-Hindi sentence pairs |
| **Pandas** | Handling dataset |
| **Regex and String Processing** | Preprocessing input text |

**4. Project Structure**

| **File** | **Purpose** |
| --- | --- |
| TTNEW.py | Main application script containing UI code, translation logic, and model loading functions. |

**5. Main Components**

**5.1. UI Components**

* Page setup with **Streamlit config**.
* **Sidebar**: Translation direction, method selection, sample dataset view.
* **Main Page**: Input text area, translation button, and translation output box.

**5.2. Core Functions**

* preprocess\_text(text, is\_hindi=False): Prepares the text by spacing out punctuation marks.
* load\_hf\_dataset(): Loads and samples from the cfilt/iitb-english-hindi dataset.
* load\_hf\_translator(model\_name): Loads Hugging Face translation pipelines.
* translate\_hindi\_to\_english\_dict(hindi\_text, translation\_dict, df): Dictionary-based Hindi ➔ English translation.
* translate\_english\_to\_hindi\_dict(english\_text, translation\_dict, df): Dictionary-based English ➔ Hindi translation.
* main(): Sets up the Streamlit page, handles user inputs, and performs translation.

**6. How It Works**

1. **User selects** translation direction and method.
2. **User inputs** Hindi or English text.
3. **On clicking Translate**:
   * If **Neural Model** is selected:
     + Uses Hugging Face pre-trained models.
   * If **Basic Lookup** is selected:
     + Looks up the dictionary for sentence match or performs word-by-word mapping.
4. **Translated output** is displayed on the screen.

**7. Requirements**

pip install streamlit pandas datasets transformers pytorch tensorflow

**8. How to Run**

streamlit run TTNEW.py

**9. Error Handling & Fixes**

**9.1 PyTorch Dependency Errors**

**Issue**: The transformers library requires PyTorch, which was missing or not installed correctly. **Fix**: Installed PyTorch with:

pip install torch

Or followed the [official installation instructions](https://pytorch.org/get-started/locally/) based on the system configuration.

**9.2 Streamlit Backend Error**

**Issue**: Occasional StreamlitBackendError due to improper usage or corrupted cache. **Fix**: Resolved by clearing the Streamlit cache:

streamlit cache clear

And ensured correct usage of @st.cache\_resource instead of deprecated decorators.

**9.3 Overfitting Errors Due to Dataset**

**Issue**: The full dataset led to overfitting or large memory usage. **Fix**: Sampled the dataset with a limit of 20,000 rows to reduce memory usage and prevent overfitting:

sample\_size = min(len(df), 20000)

sampled\_df = df.sample(n=sample\_size, random\_state=42)

**9.4 Streamlit label\_visibility Error**

**Issue**: Older versions of Streamlit didn’t support the label\_visibility parameter. **Fix**: Updated Streamlit to the latest version:

pip install --upgrade streamlit

visibility = yes/True

**9.5 Epoch Training Errors**

**Issue**: Although the app doesn’t train models during runtime, attempted model training caused instability due to improper batching or parameter selection. **Fix**: Skipped training in production app. Relied on pre-trained models only. Ensured max token limit was respected:

results = translator(input\_text, max\_length=512)

**10. Acknowledgments**

* **Helsinki-NLP** for open-sourcing the translation models.
* **CFILT-IITB** and **Hugging Face** for providing the bilingual dataset.