Language Detector and Translator

**Submitted by:**

**Kapil Raghav(102267004)**

**Agrani Shukla(102217239)**

**BE Third Year CSE**

Submitted to:

Dr. Anjula Mehto

Assistant Professor



**Computer Science and Engineering Department**

**Thapar Institute of Engineering and Technology, Patiala**

**November 2024**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S. No | Topic | Page No. |
| 1 | Introduction or Project Overview | 3 |
| 2 | Problem Statement | 3 |
| 3 | Overview of the Dataset used | 4 |
| 4 | Project workflow | 5 |
| 5 | Results | 7 |
| 6 | Conclusion | 9 |

**Project Overview**

This project focuses on developing a robust and scalable language detection and translation system, combining advanced machine learning techniques with pre-trained neural models. The primary aim is to create a pipeline capable of accurately identifying the language of a given text and translating it into a target language, ensuring high fidelity in multilingual communication.

**Problem Statement**

**Objective:** To develop an efficient system capable of identifying the language of a given text and translating it into a target language using machine learning and pre-trained models. The system should be accurate, scalable, and user-friendly.

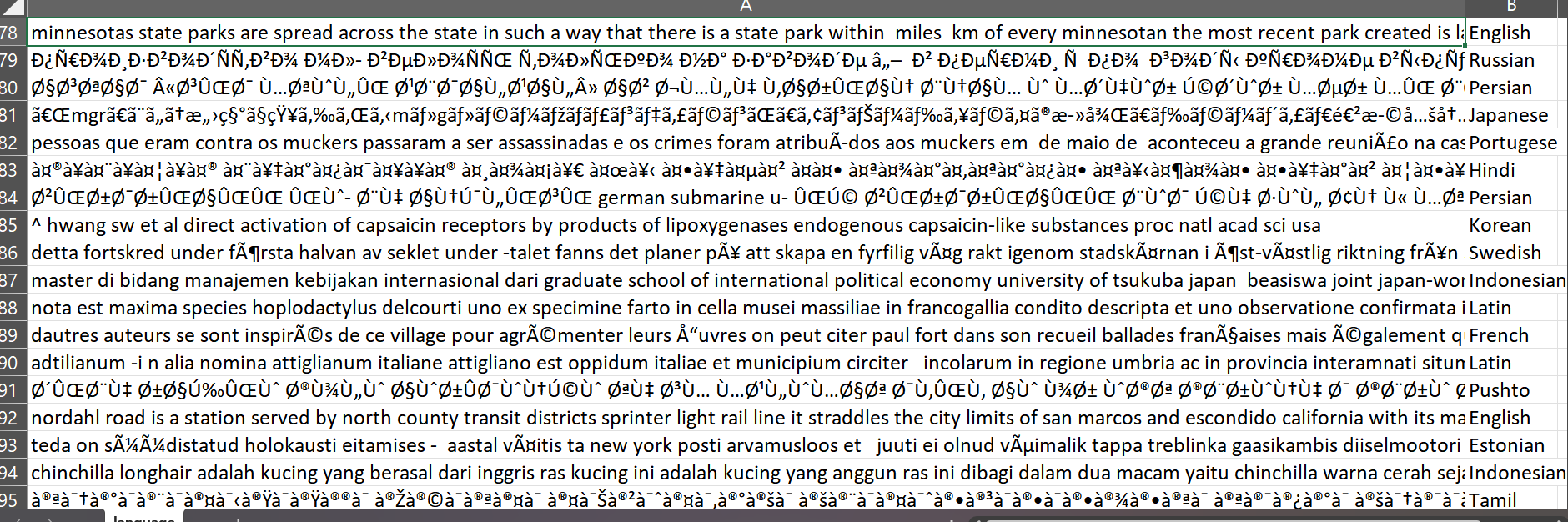
Challenges Addressed:

1. **Language Identification:**
   * Automatically determine the language of an input text from a predefined set of languages using a machine learning model.
   * Ensure robust detection across diverse and noisy input text formats.
2. **Language Translation:**
   * Translate the detected language into a user-specified target language using pre-trained neural machine translation models.
   * Handle multilingual scenarios with high fidelity and accuracy.

**Overview of the Dataset used**

The dataset comprises 22,000 entries with two key columns: Text and language. The Text column contains samples of textual data in multiple languages, while the language column provides the corresponding language labels for each entry. Both columns are complete, with no missing values, ensuring the data is reliable for analysis. The dataset represents a diverse set of languages, including Estonian, Swedish, Thai, Tamil, and Dutch, making it ideal for multilingual applications.

This dataset has been utilized to build and train a machine learning model for language prediction. By analyzing the patterns in the textual data, the model can accurately identify the language of any given text input. Additionally, the language detection model has been integrated into a translation workflow, where the detected language is used to select appropriate translation models, enabling seamless conversion of text from one language to another.



**Project Workflow**

**1. Language Detection**

Steps:

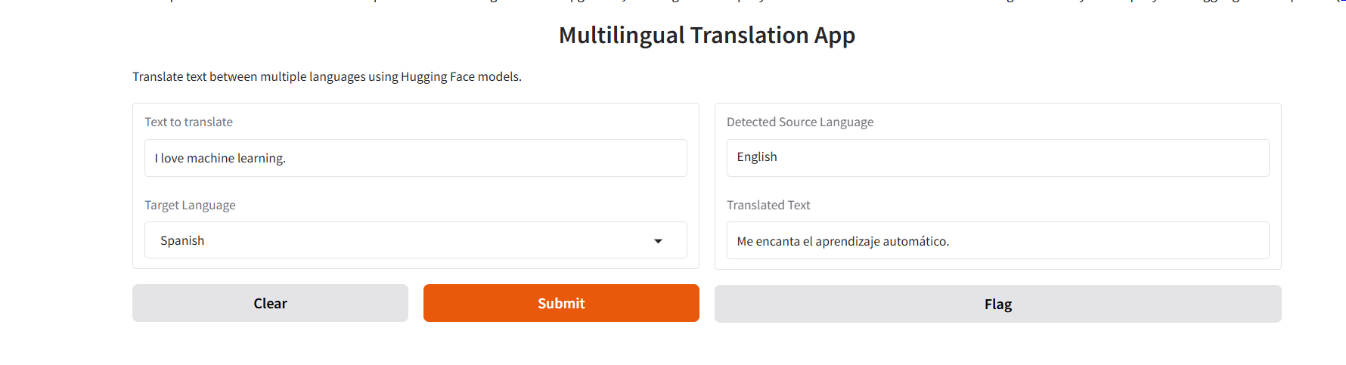
1. Dataset Preparation:
   * A CSV file (language.csv) is read into a DataFrame using pandas. This file contains textual data (Text) and corresponding language labels (language).
2. Feature Engineering:
   * The Text column is transformed into numerical data using CountVectorizer, a method that converts text into a sparse matrix of token counts.
3. Train-Test Split:
   * The dataset is split into training and testing subsets to build and evaluate the language detection model.
4. Model Training:
   * A MultinomialNB (Naive Bayes) classifier is trained on the vectorized data to predict the language.
5. Language Prediction:
   * A function detect\_language takes a text input, vectorizes it, and predicts its language using the trained model.

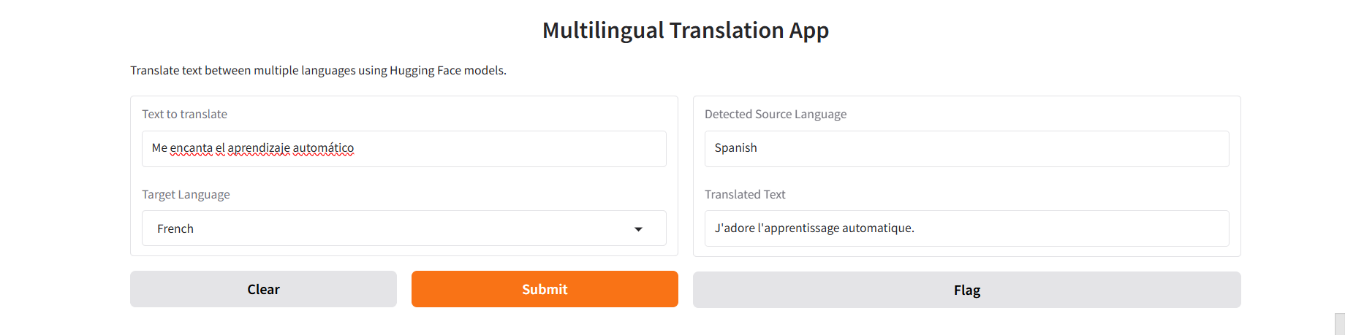
**2. Translation Model**

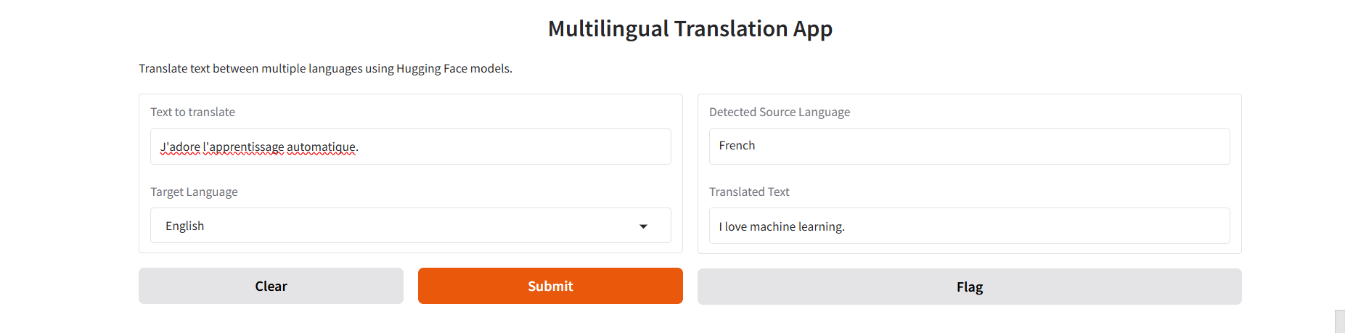
Steps:

1. Libraries Used:
   * transformers: Implements the MarianMTModel and tokenizer for machine translation.
   * gradio: Provides a web interface for testing the pipeline.
2. Model Selection:
   * A dictionary maps language pairs to pre-trained models from the Helsinki-NLP MarianMT library.
3. Translation Workflow:
   * The tokenizer processes input text for the MarianMT model.
   * The model generates translations, which are decoded back into text.
4. Integration with Gradio:
   * A Gradio app allows users to input text, select source and target languages, and view translated results interactively.

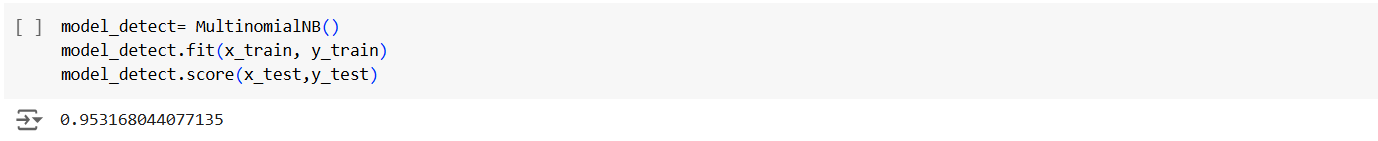
**Results**

****

****

****

Interactive Interface of the Multilingual Translation App

****

Accuracy Score of the Language Detection Model

**Conclusion**

The language detection and translation model discussed provides a robust solution for multilingual text processing. The language detection component leverages machine learning, using techniques like vectorization and Naive Bayes classification, to accurately predict the language of input text. This serves as the foundation for the translation workflow, which utilizes pre-trained MarianMT models to convert text from the detected language into a target language.

The integration of these components ensures a seamless pipeline capable of handling diverse text inputs and multilingual communication needs. This system demonstrates the power of combining machine learning and natural language processing for real-world applications like content localization, global customer support, and cross-language collaboration. With high accuracy in detection and translation, it lays the groundwork for scalable, efficient, and user-friendly multilingual solutions.

**Github Repository Link: https://github.com/agrani0613s/Language-Detector-and-Translator.git**

**THANK YOU**