Language Detector and Translator

Submitted by:

Kapil Raghav(102267004)

Agrani Shukla(102217239)

BE Third Year CSE

Submitted to:

Dr. Anjula Mehto

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

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.

	A	B
78	minnesotas state parks are spread across the state in such a way that there is a state park within miles km of every minnesotan the most recent park created is l	English
79	Ð¿Ñ€Ð¾Ð¸Ð·Ð²Ð¾Ð´ÑÑ‚Ð²Ð¾ Ð¼Ð»- Ð²ÐµÐ»Ð¾ÑÑŒ Ñ‚Ð¾Ð»ÑŒÐºÐ¾ Ð½Ð° Ð·Ð°Ð²Ð¾Ð´Ðµ â„– Ð² Ð¿ÐµÑ€Ð¼Ð¸ Ñ Ð¿Ð¾ Ð³Ð¾Ð´Ñ‹ ÐºÑ€Ð¾Ð¼Ðµ Ð²Ñ‹Ð¿Ñƒ	Russian
80	Ø§Ø³ØªØ§Ø¯ Â«Ø³ÛŒØ¯ Ù…ØªÙˆÙ„ÛŒ Ø¹Ø¨Ø¯Ø§Ù„Ø¹Ø§Ù„Â» Ø§Ø² Ø¬Ù…Ù„Ù‡ Ù‚Ø§Ø±ÛŒØ§Ù† Ø¨Ù†Ø§Ù… Ùˆ Ù…Ø´Ù‡ÙˆØ± Ú©Ø´ÙˆØ± Ù…ØµØ± Ù…ÛŒ Ø¨	Persian
81	ã€Œmgrã€ã¨ã„ã†æ„›ç§°ã§çŸ¥ã‚‰ã‚Œã‚‹mãƒ»gãƒ»ãƒ©ãƒ¼ãƒžãƒãƒ£ãƒ³ãƒ‡ã‚£ãƒ©ãƒ³ã€ã‚¢ãƒ³ãƒŠãƒ¼ãƒ‰ã‚¥ãƒ©ã‚¤ã®æ-»å¾Œã€ãƒ‰ãƒ©ãƒ¼ãƒ´ã‚£ãƒ€é€²æ-©å…šå†..	Japanese
82	pessoas que eram contra os muckers passaram a ser assassinadas e os crimes foram atribuÃdos aos muckers em de maio de aconteceu a grande reuniÃ£o na cas	Portugese
83	à¤®à¥à¤¨à¥à¤¦à¥à¤® à¤¨à¥‡à¤°à¤¿à¤¯à¤¥à¥à¤® à¤¸à¤¾à¤¡à¥€ à¤œà¥‹ à¤•à¥‡à¤µà¤² à¤à¤• à¤ªà¤¾à¤°à¤¸à¤ªà¤°à¤¿à¤• à¤ªà¥‹à¤¶à¤¾à¤• à¤•à¥‡à¤°à¤² à¤¦à¤•à¥	Hindi
84	Ø²ÛŒØ±Ø¯Ø±ÛŒØ§ÛŒÛŒ ÛŒÙˆ- Ø¨Ù‡ Ø§Ù†Ú¯Ù„ÛŒØ³ÛŒ german submarine u- ÛŒÚ© Ø²ÛŒØ±Ø¯Ø±ÛŒØ§ÛŒÛŒ Ø¨ÙˆØ¯ Ú©Ù‡ Ø·ÙˆÙ„ Ø¢Ù† Ù« Ù…Øª	Persian
85	^ hwang sw et al direct activation of capsaicin receptors by products of lipoxygenases endogenous capsaicin-like substances proc natl acad sci usa	Korean
86	detta fortskred under fÃ¶rsta halvan av seklet under -talet fanns det planer pÃ¥ att skapa en fyrfilig vÃ¤g rakt igenom stadskÃ¤rnan i Ã¶st-vÃ¤stlig riktning frÃ¥n	Swedish
87	master di bidang manajemen kebijakan internasional dari graduate school of international political economy university of tsukuba japan beasiswa joint japan-wo	Indonesian
88	nota est maxima species hoplodactylus delcourti uno ex specimine farto in cella musei massiliae in francogallia condito descripta et uno observatione confirmata i	Latin
89	dautres auteurs se sont inspirÃ©s de ce village pour agrÃ©menter leurs Å“uvres on peut citer paul fort dans son recueil ballades franÃ§aises mais Ã©galement q	French
90	adtilianum -i n alia nomina attiglianum italiane attigliano est oppidum italiae et municipium circiter incolarum in regione umbria ac in provincia interamnati situn	Latin
91	Ø´ÛŒØ¨Ù‡ Ø±Ø§Ú‰ÛŒÙˆ Ø®Ù¾Ù„Ùˆ Ø§ÙˆØ±ÛØ¯ÙˆÙ†Ú©Ùˆ ØªÙ‡ Ø³Ù… Ù…Ø¹Ù„ÙˆÙ…Ø§Øª Ø¯Ù‚ÛŒÙ‚ Ø§Ùˆ Ù¾Ø± ÙˆØ®Øª Ø®Ø¨Ø±ÙˆÙ†Ù‡ Ø¯ Ø®Ø¨Ø±Ùˆ Ø	Pushto
92	nordahl road is a station served by north county transit districts sprinter light rail line it straddles the city limits of san marcos and escondido california with its ma	English
93	teda on sÃ¼Ã¼distatud holokausti eitamises - aastal vÃ¤itis ta new york posti arvamusloos et juuti ei olnud vÃµimalik tappa treblinka gaasikambis diiselmootori	Estonian
94	chinchilla longhair adalah kucing yang berasal dari inggris ras kucing ini adalah kucing yang anggun ras ini dibagi dalam dua macam yaitu chinchilla warna cerah sej	Indonesian
95	à®ªà¯†à®°à¯à®¨à¯à®¤à¯‹à®Ÿà¯à®Ÿà®®à¯ à®Žà®©à¯à®ªà®¤à¯ à®¤à¯Šà®²à¯ˆà®¤à¯‚à®°à®šà¯ à®šà®¨à¯à®¤à¯ˆà®•à®³à¯à®•à¯à®•à®¾à®•à®ªà¯ à®ªà®¯à®¿à®°à¯ à®šà¯†à®¯à¯	Tamil

Project Workflow

1. Language Detection

Steps:

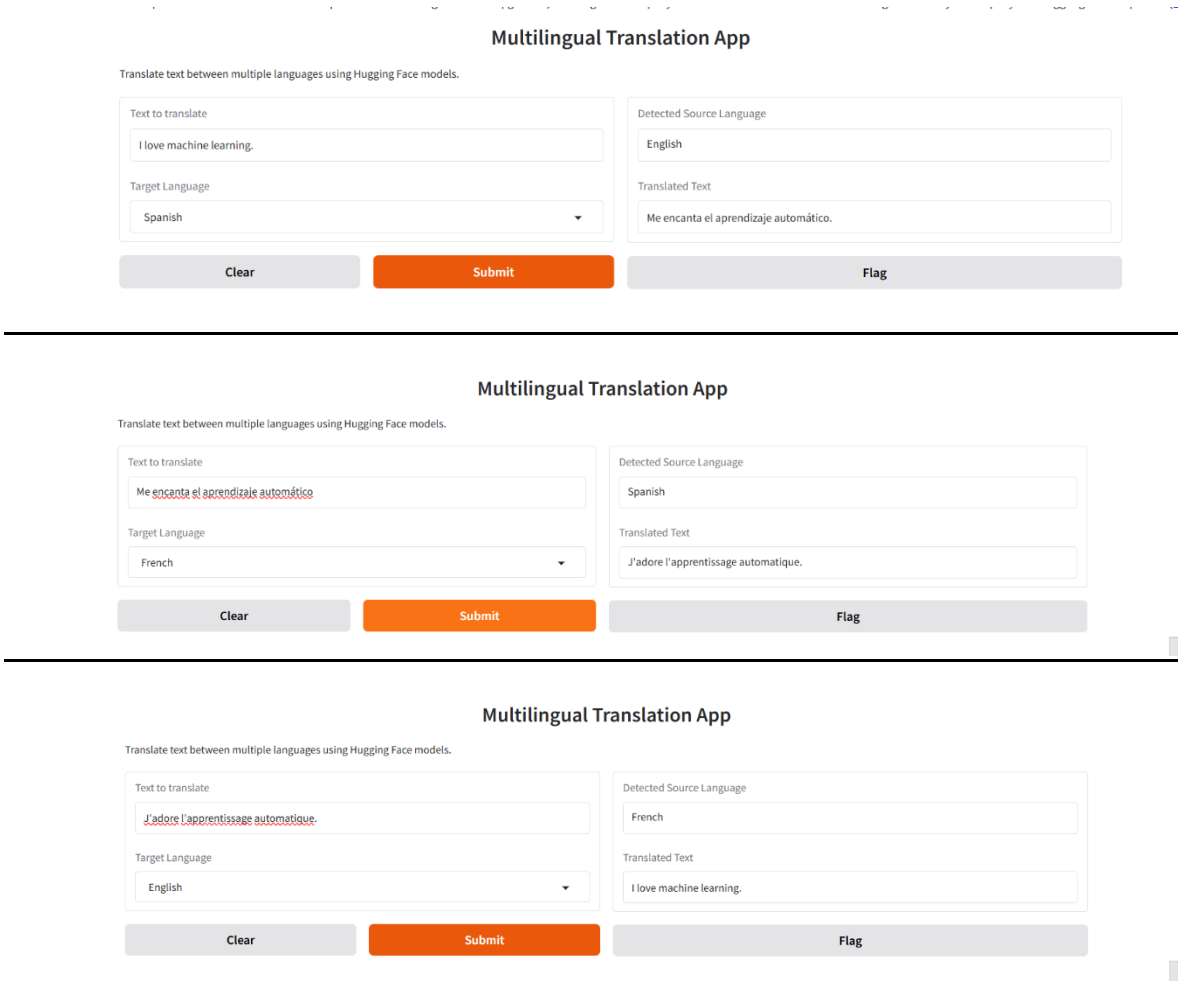
1. Dataset Preparation:
    - A CSV file (language.csv) is read into a DataFrame using pandas. This file contains textual data (Text) and corresponding language labels (language).
2. Feature Engineering:
    - The Text column is transformed into numerical data using CountVectorizer, a method that converts 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

```
[ ] model_detect= MultinomialNB()
    model_detect.fit(x_train, y_train)
    model_detect.score(x_test,y_test)
```

```
0.953168044077135
```

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