Language Detector

(regional languages)

This project is made to recognize regional languages like Hindi, Kannada, Bengali, Marathi and Telugu which are in turn written in english which makes it more challenging.

For this project we have use 2 datasets

- 1 contained Hindi , Kannada, Bengali, Marathi, Telugu and Tamil
  https://www.kaggle.com/datasets/parthplc/hindi-to-hinglish
  It was almost in the required format and just had to remove 1 column but the bad thing about this dataset was it had wrong data for telugu and tamil. Hence had to remove rows with language Telugu and Tamil from this dataset.

- As we removed languages we needed to fill the space hence had to fill it up. The 2nd dataset
  https://github.com/SunilGundapu/Word-Level-Language-Identification-in-English-Telugu-Code-Mixed-Data/blob/master/Updated%20Code%20With%20Simple%20ML%20Models/LIDataset/CodemixedShuffle.txt

  Is in txt format and also in a very different format than the required one hence was cleaned, converted to required format split for train and test train was concatenated with the train of 1st dataset and test was concatenated with valid of 1st dataset.

- Once the data was cleaned and made into required format, the model was trained using fasttext.
  https://github.com/facebookresearch/fastText
  A very good model for classification this model works like a mixture of Skipgram and CBOW model.

- At the end after training testing is done the model is saved as yo.bin Then can be used anytime by using ft.load_model('path') and then predict language of text using model.predict('text')

- A web application was made using the 'pywebio' library for this implementation which takes text entry as input and displays the predicted language. (currently on local host)

We also tried transliteration just to convert english hindi to hindi script.
Its own readme is attached in the transliteration folder zip