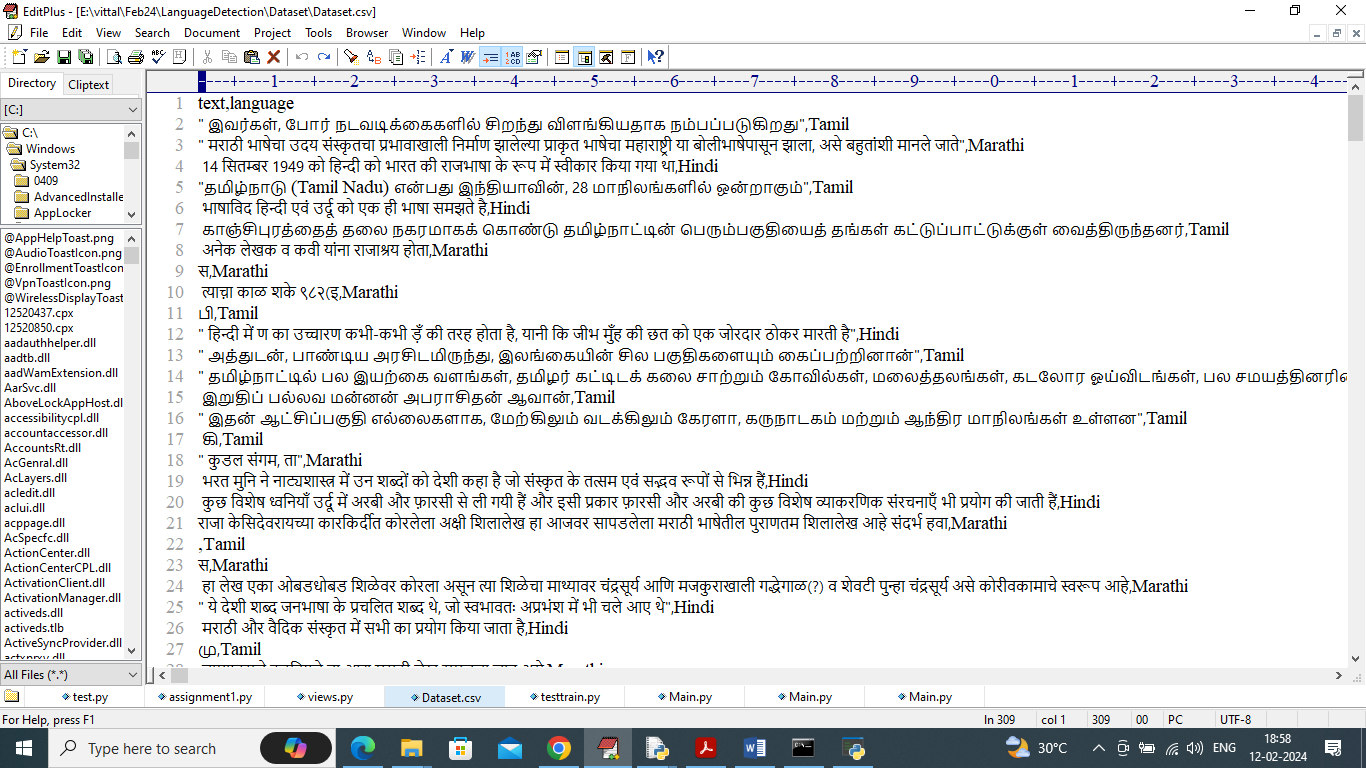
Language Identification for Multilingual Machine Translation

In this project we have employed NGRAM and Machine learning algorithms to identify language names from given text. To evaluate performance we have utilized various machine learning algorithms such as SVM, KNN and Random Forest. Each algorithm performance is tested in terms of accuracy, precision, recall, Confusion matrix graph and FSCORE. Among all algorithms Random Forest is giving high accuracy.

To train above algorithms we have used dataset of languages such as Tamil, Hindi and Marathi and this dataset can be downloaded from below URL

<https://www.kaggle.com/datasets/sandeepbelamagi/indian-local-languages>

In below screen we are showing dataset details



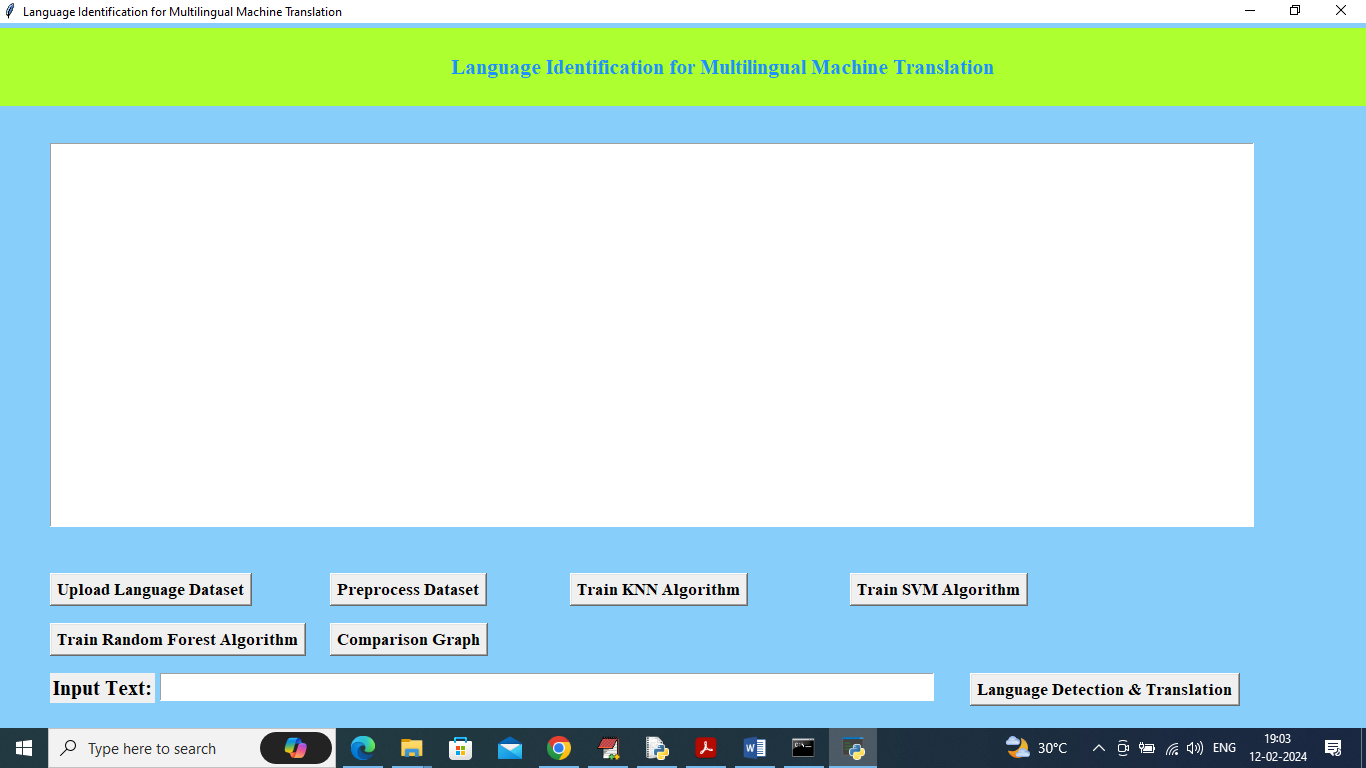
In above dataset first row contains dataset column names and remaining rows contains Text sentences and language names and by using above dataset we will train and test each algorithm performance.

To implement this project we have designed following modules

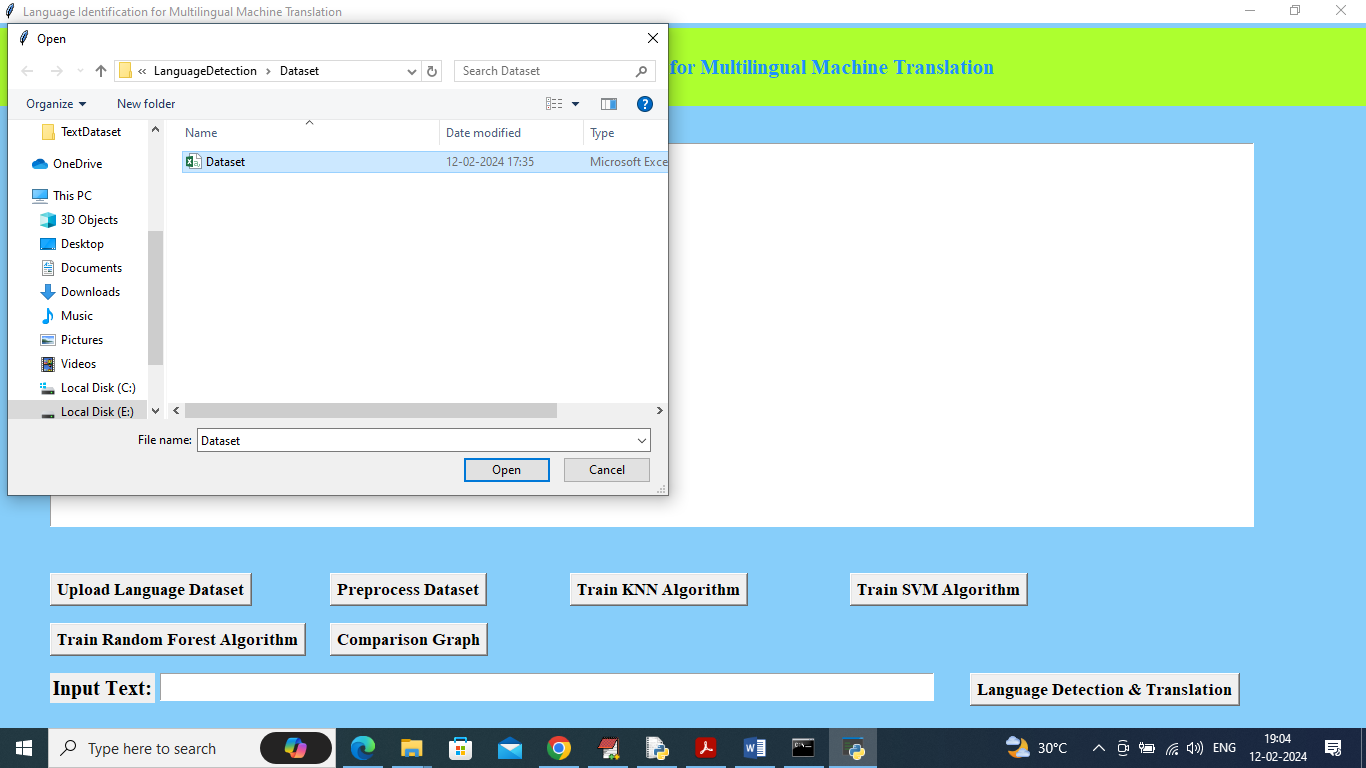
1. Upload Language Dataset: using this module we will upload dataset and then remove all missing and special symbols from dataset
2. Pre-process Dataset: using this module we will convert above process dataset into numeric vector by employing 3 NGRAMS technique and then convert entire text data into numeric vector and then split training data into train and test where application using 80% dataset for training and 20% for testing
3. Train KNN Algorithm: 80% training data will be input to KNN algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
4. Train SVM Algorithm: 80% training data will be input to SVM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
5. Train Random Forest Algorithm: 80% training data will be input to Random Forest algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
6. Comparison Graph: will plot comparison between all algorithms
7. Language Detection & Translation: here user can enter some text line and then application will predict language name and then translate that language into English using Google Translator.

SCREEN SHOTS

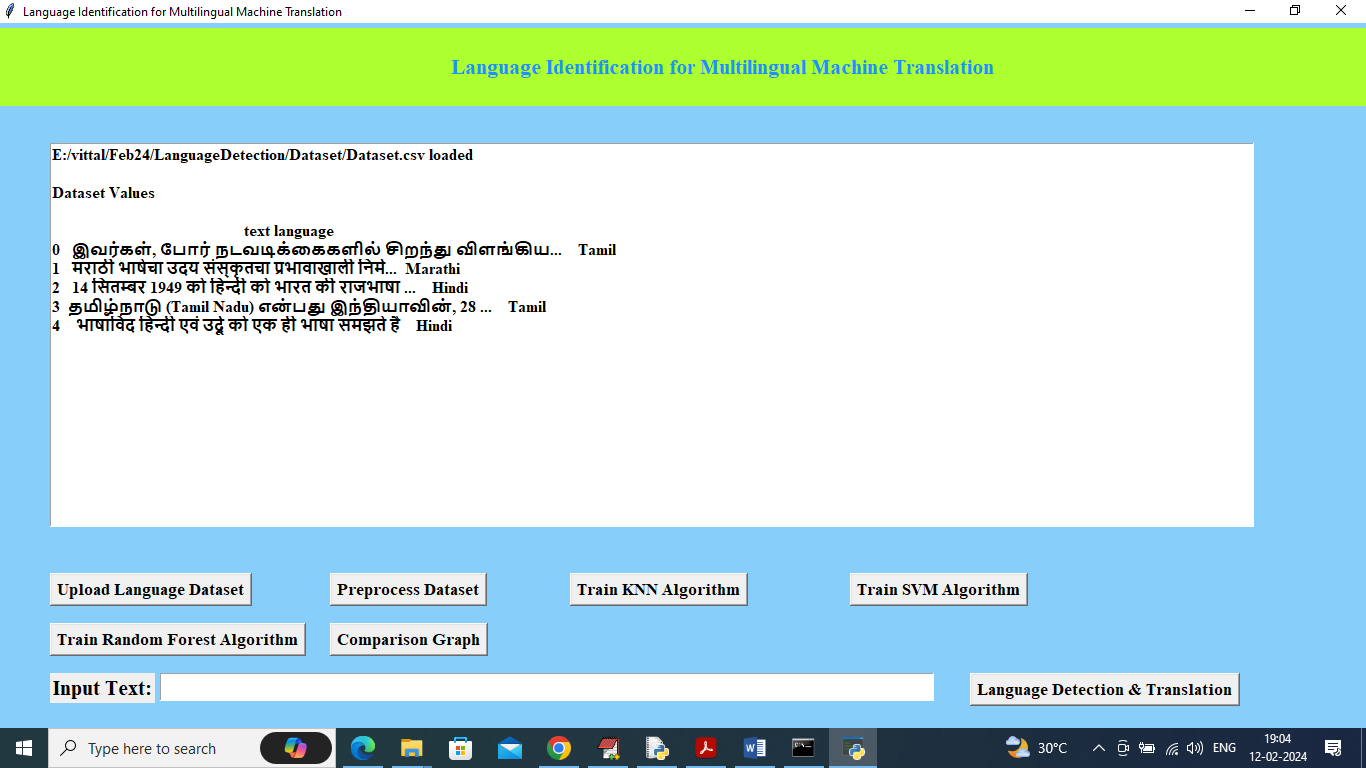
To run project double click on run.bat file to get below screen



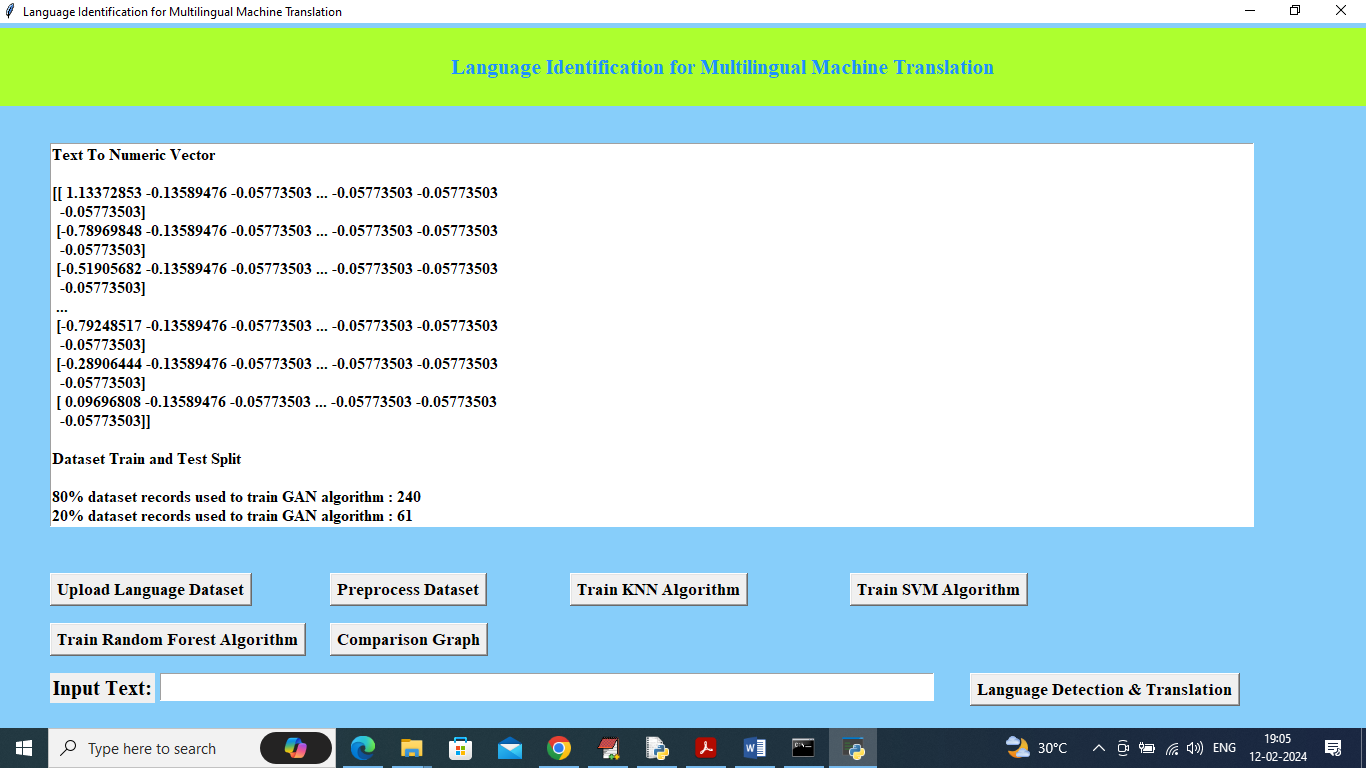
In above screen click on ‘Upload Language Dataset’ to load dataset and get below output



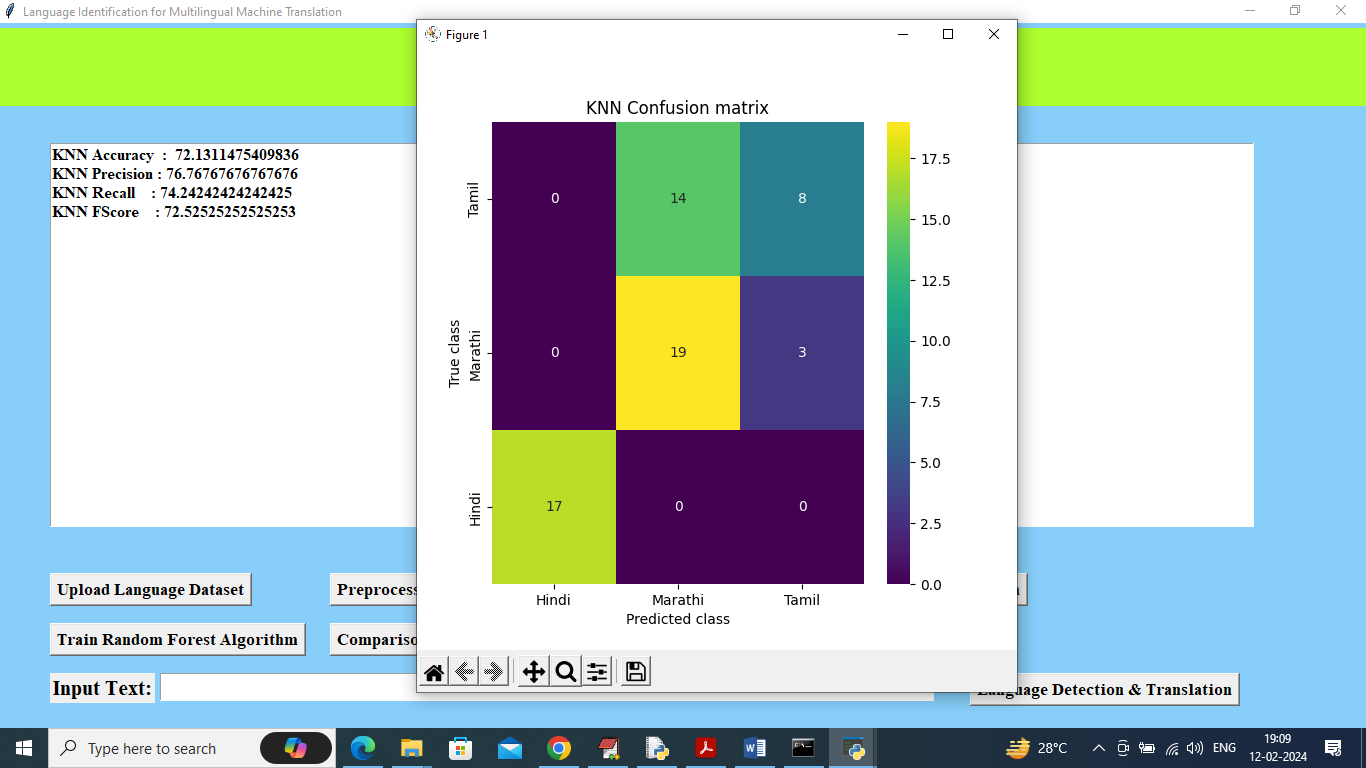
In above screen selecting and uploading dataset file and then click on ‘Open’ button to load dataset and get below page



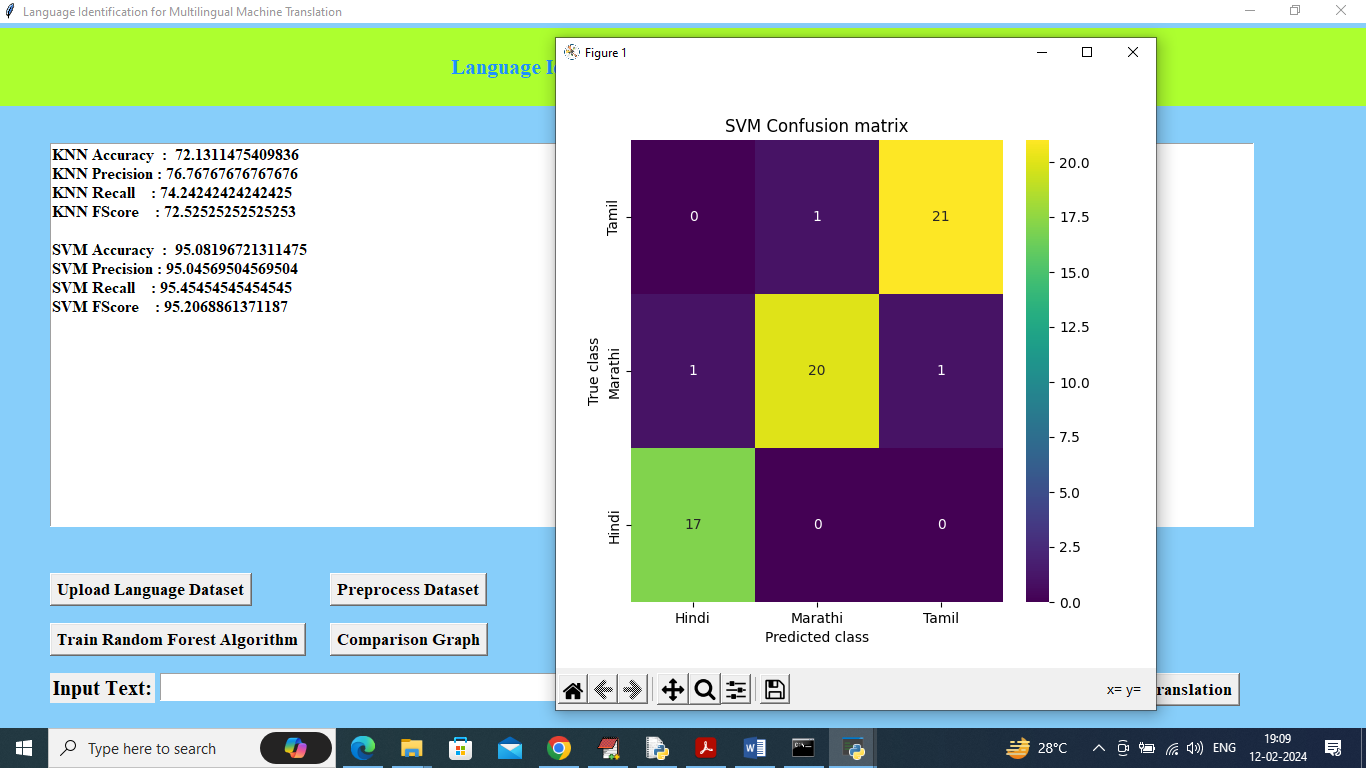
In above screen dataset loaded and now click on ‘Pre-process Dataset’ button to clean dataset and get below output



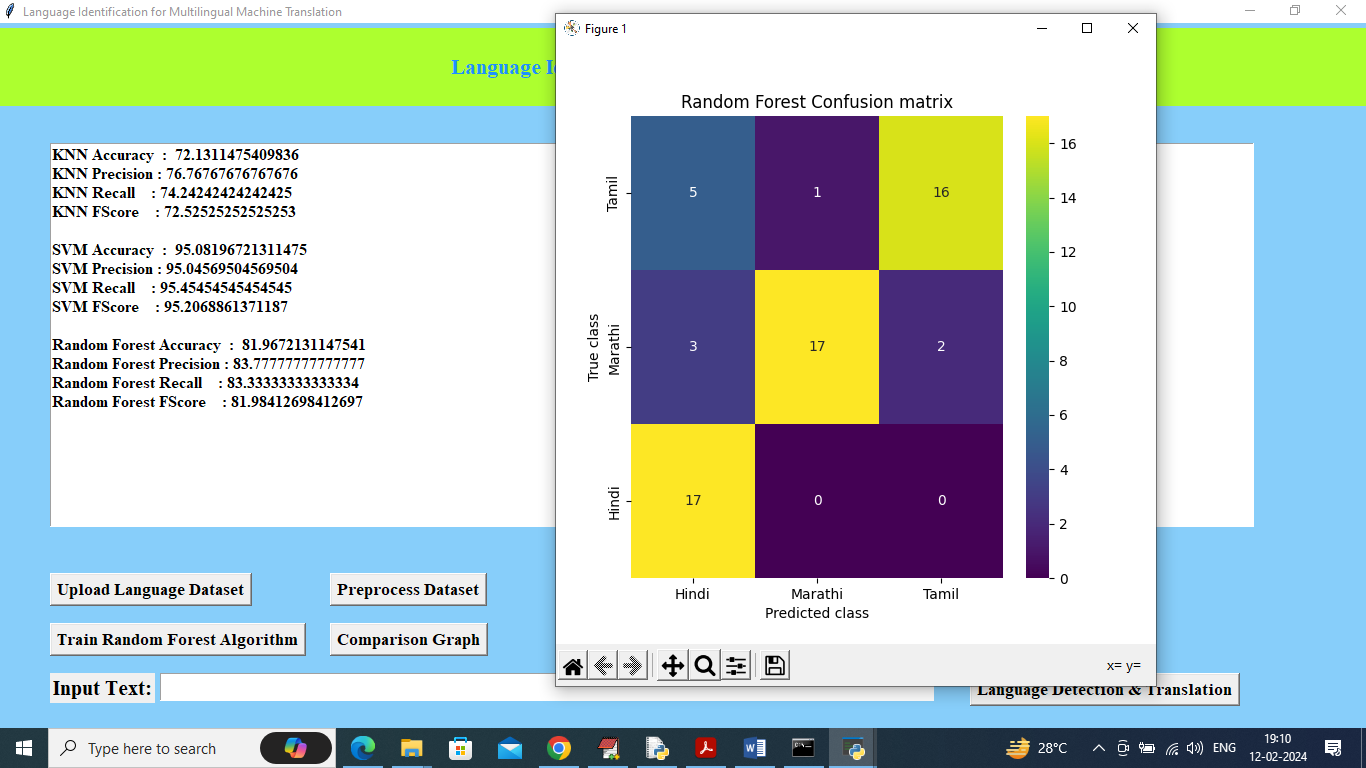
In above screen entire text data converted to numeric vector by using 3 NGRAM techniques and then can see train and test split details and now click on ‘Run KNN Algorithm’ to train KNN and get below output



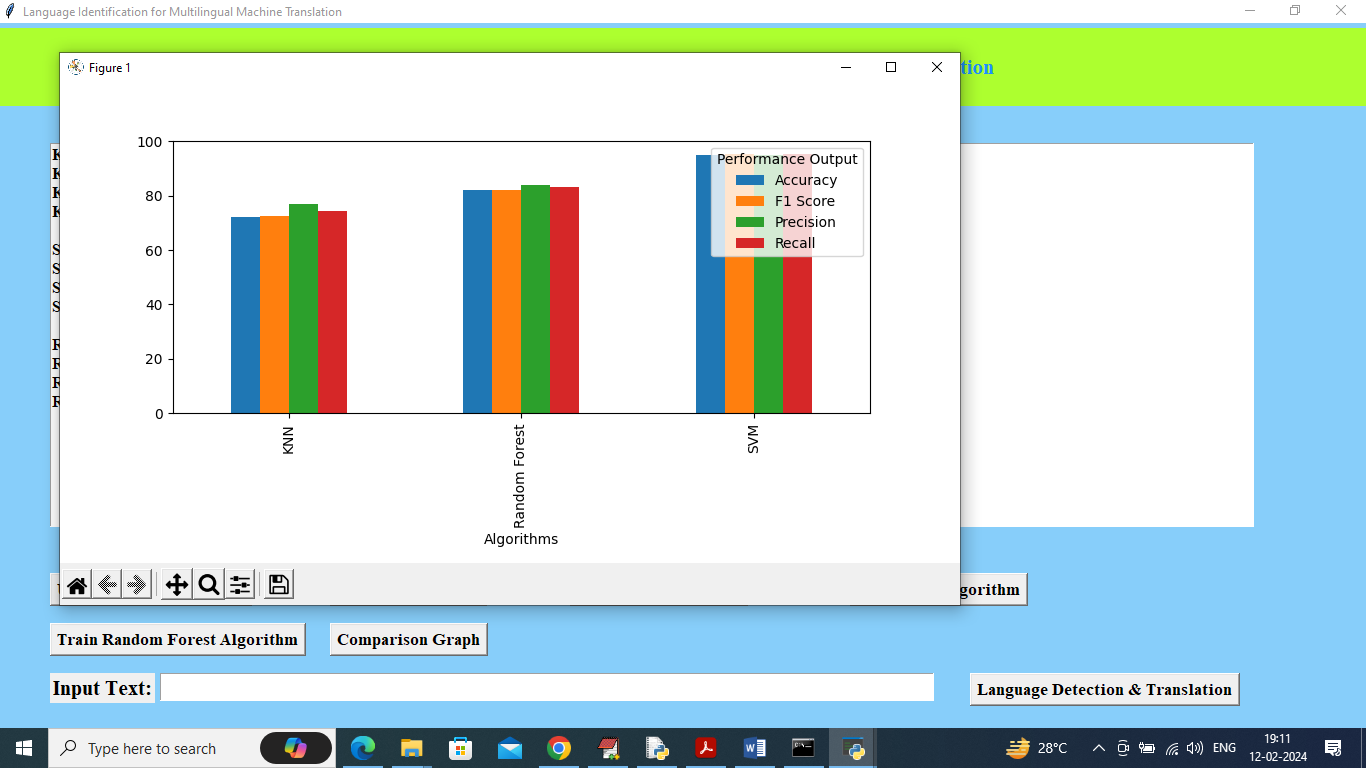
In above screen KNN training completed and it got accuracy as 72% and can see other metrics also and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and all yellow and green colour boxes in diagnol represents correct prediction count and remaining blue boxes represents incorrect prediction count and now close above graph and then click on ‘Train SVM’ button to get below output



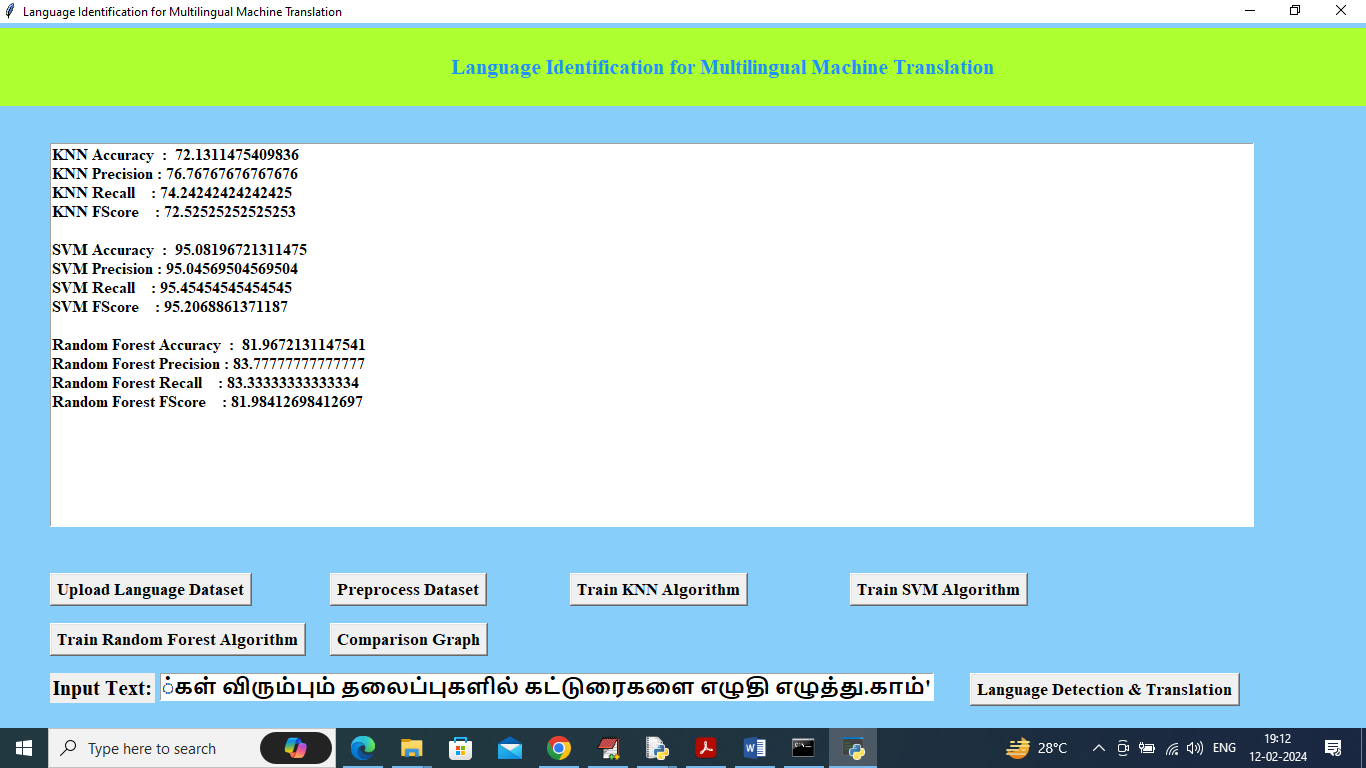
In above screen SVM got 95% accuracy and can see other metrics also and now click on ‘Train Random Forest’ to get below output



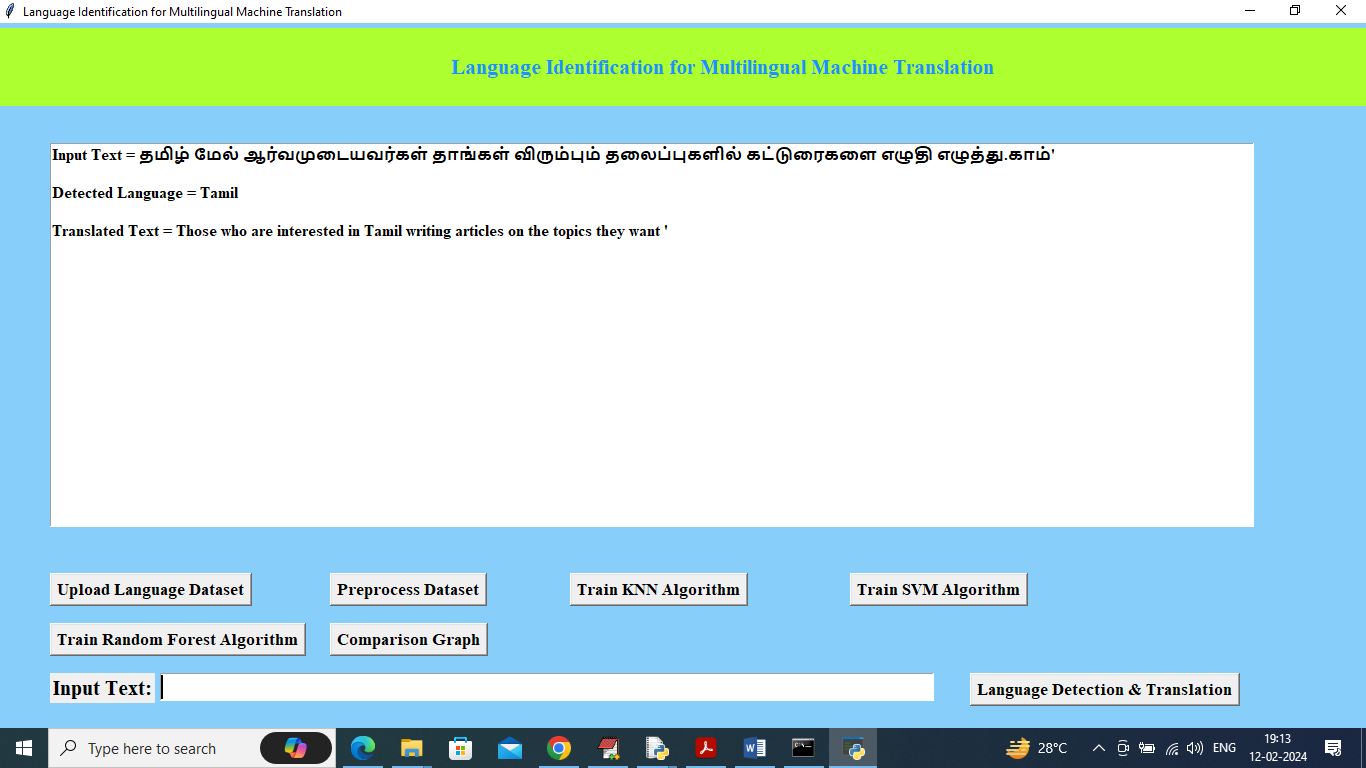
In above screen Random Forest got 81% accuracy and now click on Comparison Graph button to get below output



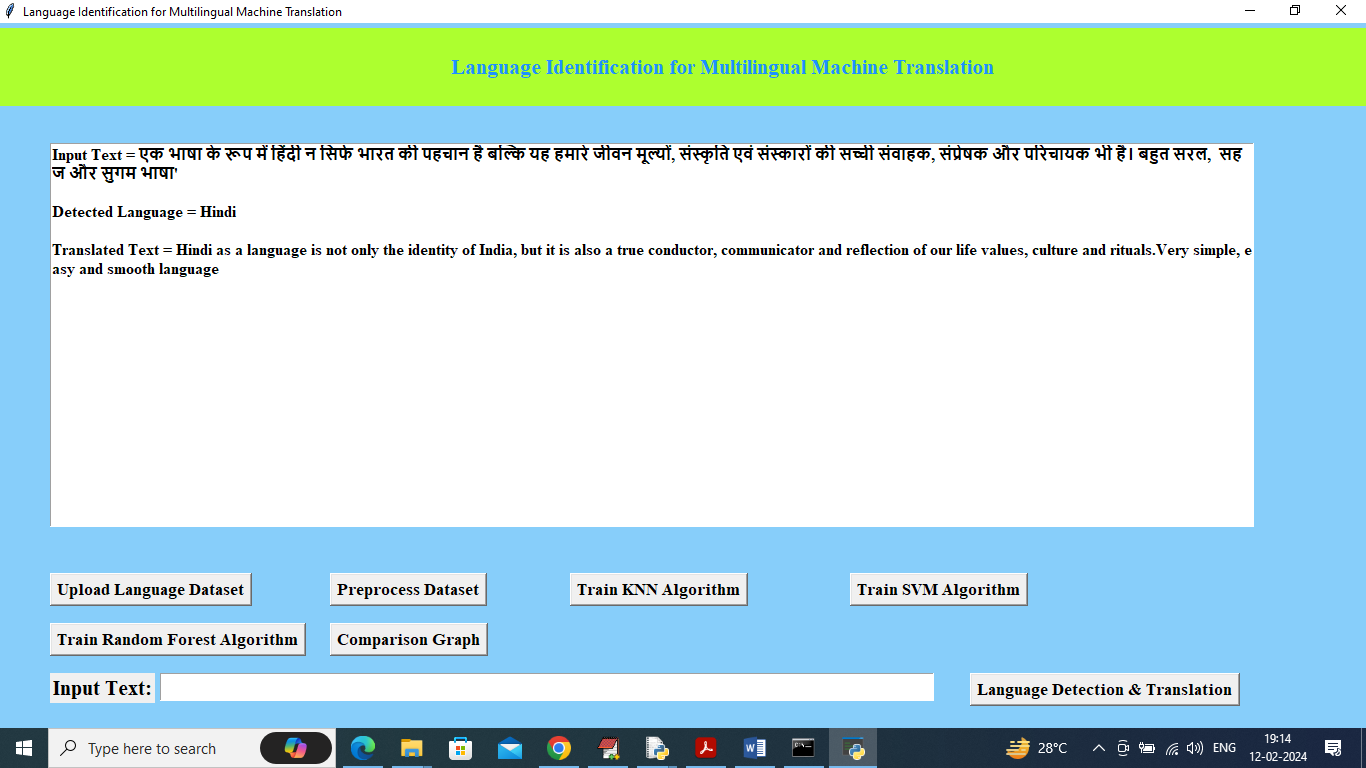
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms SVM got high accuracy and now enter some sentence in text field and then press ‘Language Detection and Translation’ button



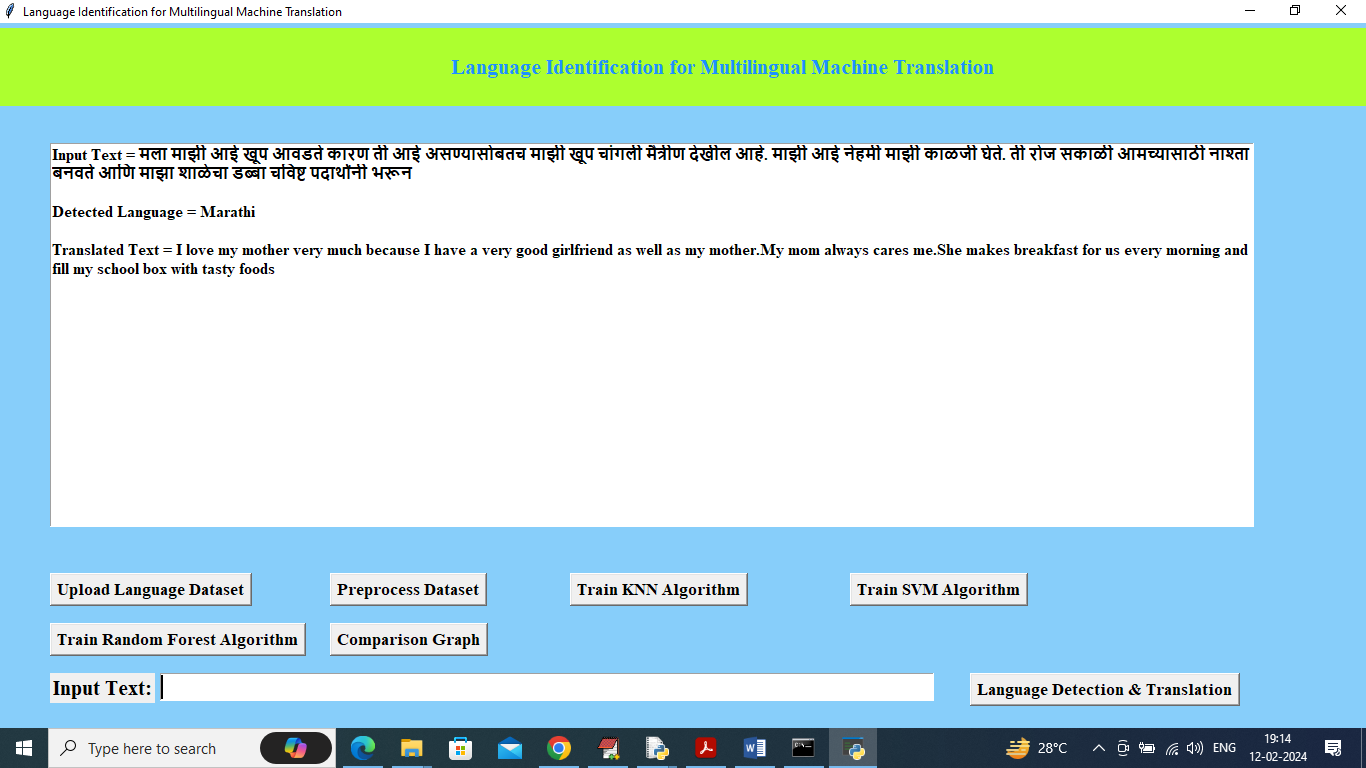
In above screen I entered some text in text field and then press Language Detect button to get below output



In above screen in text area can see Detected Language is Tamil and can see Translated text in English and below is another example



In above screen detected language is Hindi with translation



In above screen detected language is Marathi with English translation.

Similarly enter sentence in text field and get detected language and translation