Cross-Linguistic Communication – A Natural Language Processing Approach

A Capstone Project by

Varun Bodla - XI65011

Under the guidance of

Dr. Chaojie Wang

University of Maryland Baltimore County

Objective

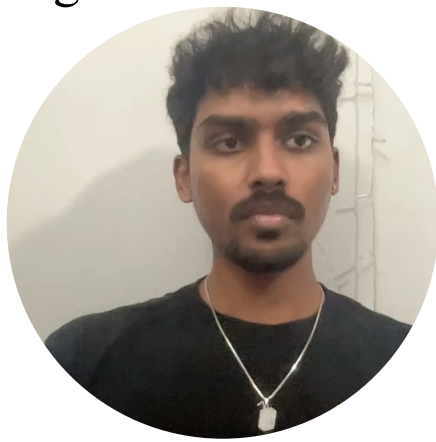
The main aim of this study is to develop language translation system leveraging deep learning techniques in NLP to enable accurate translations between Hindi to English language.

Need of this project

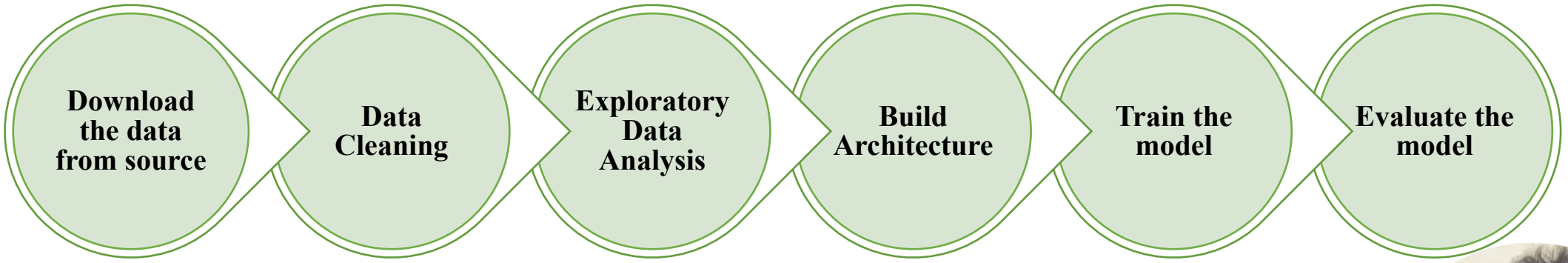
- In a globally interconnected world with diverse linguistic communities, effective communication across language barriers is essential.
- Traditional language translation methods often face challenges in capturing the contextual nuances, handling ambiguity, and providing accurate and contextually relevant translations.
- The need for a robust language translation solution is evident in various domains, including business, diplomacy, education, healthcare, and everyday communication.
- The goal is to develop a state-of-the-art language translation system that outperforms traditional methods, providing accurate, context-aware, and user-friendly translation[illegible] a diverse set of languages and applications.
- The success of this project will contribute to breaking down language barriers[illegible] global communication, and promoting inclusivity in an increasingly interconnec[illegible]

Dataset

- The dataset for this study would be The IIT Bombay English-Hindi Corpus compiled by the Indian Institute of Technology Bombay (IIT Bombay) for research purposes. This corpus consists of parallel text data in both English and Hindi languages.
- The dataset is of the size 99.7 MB
- This dataset contains pairs of sentences in English and Hindi, where each English sentence corresponds to its equivalent in Hindi. The corpus likely encompasses a diverse range of topics and genres to capture the variations in language usage across different contexts. It may include texts from various sources such as news articles, literature, websites, and possibly even user-generated content.



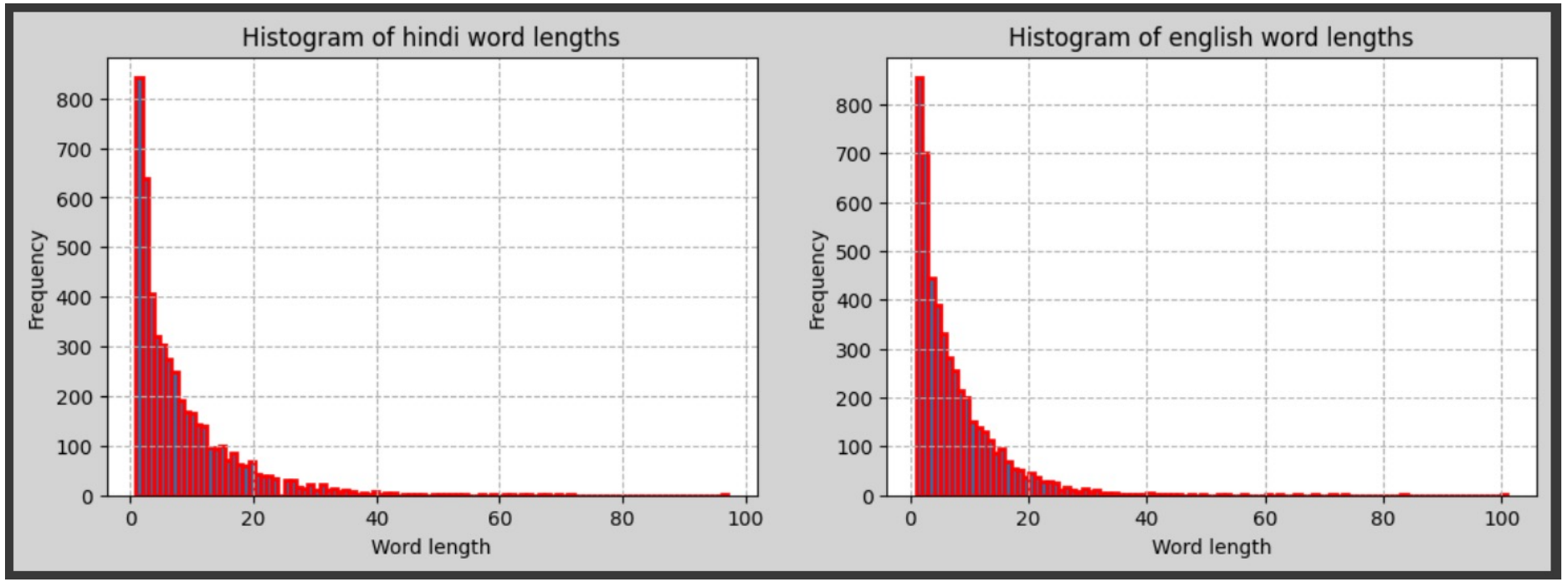
Methodology



Data cleaning

- Originally, the dataset has 1.65 million datapoints in both English and Hindi languages. In order to simplify the study considering the computation resources, 5000 points have been randomly sampled from the original dataset.
- The following text cleaning steps have been performed.
  - Apply decontractions for English language text. For example, can’t to cannot and wouldn’t to would not.
  - Lower case the text.
  - Remove all the non-alphabetic characters from the text string leaving only letters.
  - For Hindi language text, extracting only Hindi word text.
  - Removing single letter words
  - Removing stop words from the English text.
  - Dropping rows with empty strings in the Hindi and English columns.
- The effective shape of the dataset has become 4859 datapoints.

Exploratory data analysis



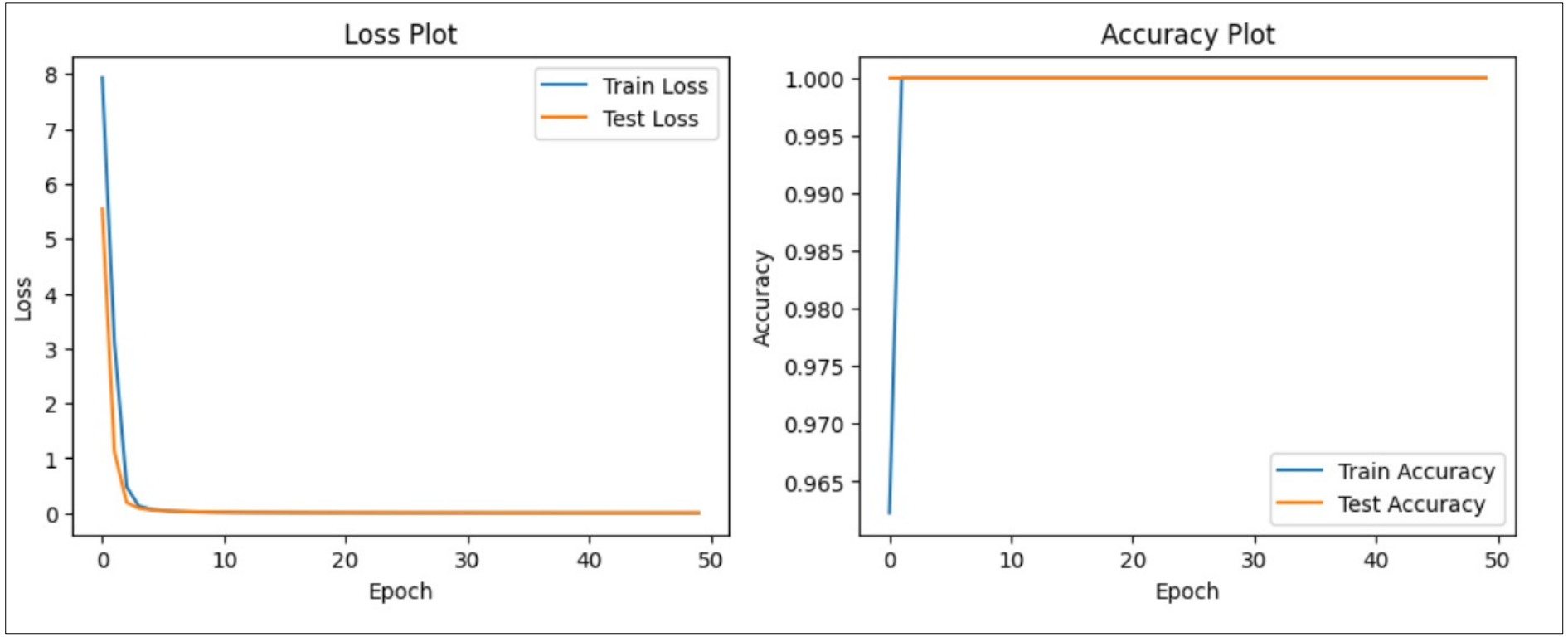
It can be observed from the above histograms that the distribution of word lengths for both Hindi and English are skewed towards right. It can be inferred that there are more sentences with less number of words in text in both English and Hindi languages for the given data.

Model architecture

```
Model: "encoder_decoder_1"
_________________________________________________________________
 Layer (type)                                Output Shape                                Param #
=================================================================
 encoder_1 (Encoder)                         multiple                                    1294224
|-----------------------------------------------------------------------------------------------|
| Encoder_Embedding_Layer (Embedding)        multiple                                    1270800     |
|                                                                                                    |
| Encoder_LSTM (LSTM)                        multiple                                    23424       |
-------------------------------------------------------------------------------------------------
 decoder_1 (Decoder)                         multiple                                    1447824
|-----------------------------------------------------------------------------------------------|
|                                                                                                    |
| embedding_layer_decoder (Embedding)        multiple                                    1424400     |
|                                                                                                    |
| Decoder_LSTM (LSTM)                        multiple                                    23424       |
-------------------------------------------------------------------------------------------------
 dense_1 (Dense)                             multiple                                    313335

=================================================================
Total params: 3055383 (11.66 MB)
Trainable params: 3055383 (11.66 MB)
Non-trainable params: 0 (0.00 Byte)
```

Results


Loss Plot
Train Loss
Test Loss
Loss
Epoch
0
1
2
3
4
5
6
7
8
0
10
20
30
40
50
Accuracy Plot
Train Accuracy
Test Accuracy
Accuracy
Epoch
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0.965

Results

- We have achieved a test accuracy of 100% by training 3.05 million parameters for 50 epochs using Adam optimizer with an initial learning rate of 0.001 with a batch size of 64.

Conclusion

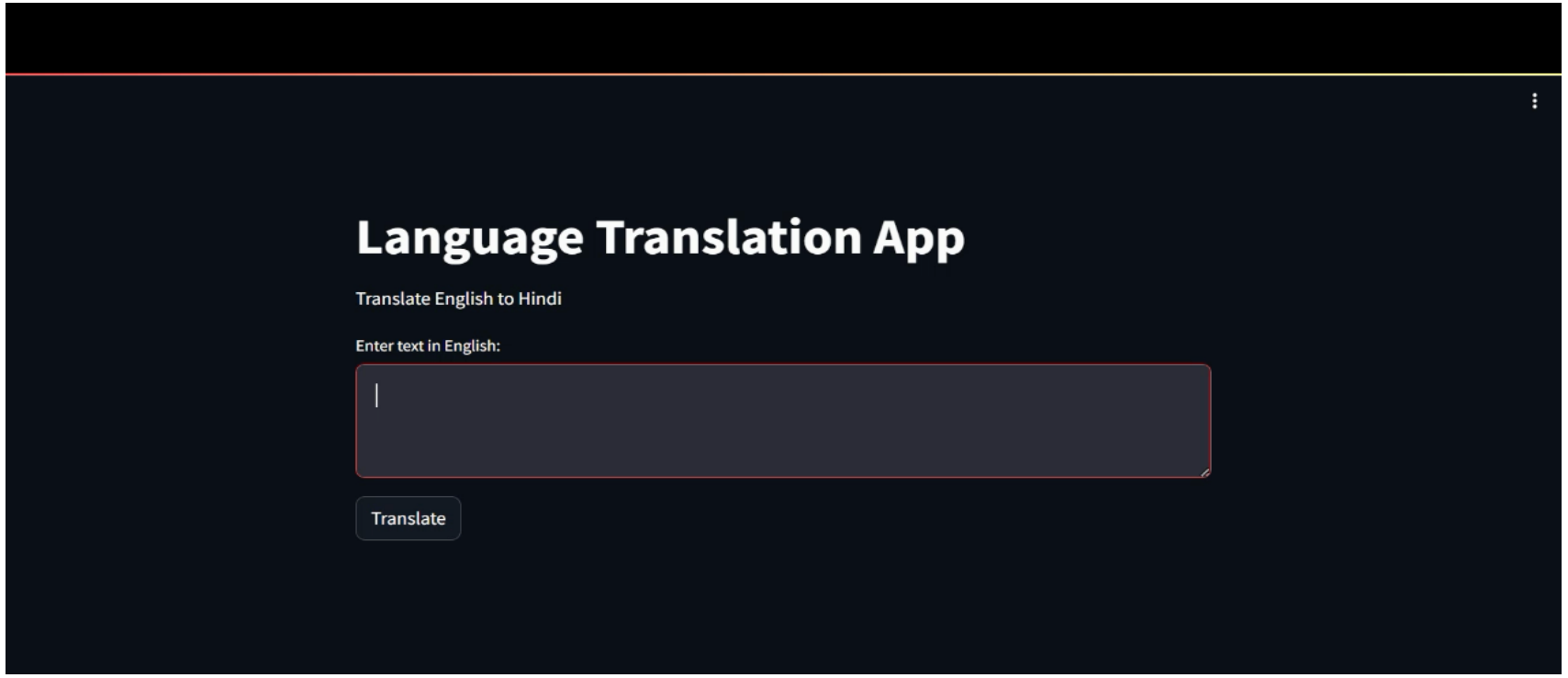
The machine translation model from English to Hindi performs exceptionally well with perfect accuracy and very low error rates on the provided data. This excellent performance offers several real-world benefits:

- Better Access - The model makes it easier for Hindi-speaking people to access information and services that were previously available only in English.
- Enhanced Communication - It improves communication in personal, professional, and business contexts, helping people understand each other better.
- Business Expansion - Companies can reach more customers in the Hindi-speaking market, driving business growth.
- Educational Support - Students and learners can access a wider range of educational materials, aiding their learning and development.
- Healthcare Improvement - Hindi-speaking patients can get accurate medical information a leading to better health outcomes.

In summary, the model's ability to accurately translate between English and Hindi ca impact various areas, making information and services more accessible and communi effective.

Deployment

The web application has been built using Streamlit and has been deployed in Streamlit Cloud.



Link - https://language-translation-application.streamlit.app/

Future Work

- Due to the limited computation resources, the model has been trained with just 5000 datapoints. The model could very well be learned with all the 1.65 million points with a good neural machine and therefore can make the model generalize well on the unseen data.
- The Attention layer can be added to the encoder decoder model to learn the long sequences effectively.