UPOS Tagging Using Sequence To Sequence Model

Introduction :

Sequence to Sequence models is a special class of Recurrent Neural Network architectures to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc. In this project, Sequence to Sequence Model is used to perform Universal Part Of Speech tagging. The data is extracted from team 1, team 2 and Gold dataset. The data in team sheets is used as a training set and Gold dataset used as a Validation set, Various Neural Network models are implemented and their performance is evaluated.

About Data :

Training Data :

Data is extracted from team 1 and team 2 from all 4 categories of corpus. The data is extracted from News, Movies, Wiki and Ebook Dataset. The sentence in the Gold dataset is excluded and manually removed. Since it is used in Validation.

Number Of Rows : 103331

Number Of Sentences : 6260

Validation Data :

Validation Data is only extracted from the Gold dataset. The Gold dataset is used as a validation Data Here.

Number Of Rows : 10317

Number Of Sentences : 707

Implementation :

Preprocess data :

Data is loaded and words are extracted with their respective POS Tags. Then, words and **UPOS tags** are converted into a list of lists. Then data is encoded to numbers with keras Tokenizer. Then the data is padded with **pre-padding of size 34**, since the longest sentence in the dataset consists of **34 token**s. Padding is used to normalize the input size.

Implementation Of Model :

Splitting of the dataset into train and validation is a part of data augmentation of the model. Gold dataset is used as test data. We did a one-hot encode for the output sequence. We used a keras embedding layer for word embedding. The maximum length of the sentence is found and passed for embedding. Chiefly, four algorithms have been implemented.The sentence is **pre-padded** with size of **29** to **normalize** the data. The size of **29** is chosen since the **longest sentence** in the data consists of **29 tokens**. They are: **LSTM, CNN, Bidirectional LSTM and Simple RNN.** For fairness sake, all these models have **embedding size** of **300**, **output size** of **19** for categories and **input shape** of **(None,29)** representing maximum length of the sentence in the dataset. After training data in each model, all the model's performance is evaluated with the gold dataset (Test Dataset) . And test accuracy is calculated for each model. The each model's performance is tabulated below :

Model Evaluation :

Table 1.1. Performance Of LSTM model

S.No	Epochs	Time (In Sec)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Testing accuracy
1	5	21 sec	0.1219	0.9663	0.2348	0.9350	0.847
2	10	35 sec	0.0426	0.9876	0.2923	0.9360	0.8478
3	15	49 sec	0.0278	0.9904	0.3093	0.9367	0.8472
4	20	64 sec	0.0249	0.9913	0.2977	0.9397	0.8488
5	25	87 sec	0.0212	0.9916	0.2993	0.9376	0.8475

Table 1.2. Performance Of CNN model

S.No	Epochs	Time (In Sec)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Testing accuracy
1	5	1 sec	0.2743	0.9087	0.3797	0.8886	0.839
2	10	1 sec	0.0525	0.9844	0.3356	0.9340	0.8477
3	15	1 sec	0.0320	0.9887	0.3494	0.9359	0.8466
4	20	2 sec	0.0297	0.9899	0.4021	0.9317	0.8464
5	25	2 sec	0.0527	0.9836	0.5739	0.9211	0.8425

Table 1.3. Performance Of Bidirectional LSTM model

S.No	Epochs	Time (In Sec)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Testing accuracy
1	5	27 sec	0.3332	0.9359	0.4127	0.9182	0.8471
2	10	53 sec	0.0560	0.9827	0.2441	0.9407	0.8498
3	15	78 sec	0.0272	0.9904	0.2646	0.9409	0.8488
4	20	102 sec	0.0185	0.9935	0.3040	0.9401	0.8486
5	25	127 sec	0.0152	0.9943	0.3005	0.9402	0.8482

Table 1.4. Performance Of Simple RNN model

S.No	Epochs	Time (In Sec)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Testing accuracy
1	5	16 sec	0.1208	0.9677	0.2201	0.9381	0.8494
2	10	33 sec	0.0370	0.9883	0.2344	0.9414	0.8489
3	15	47 sec	0.0263	0.9908	0.2684	0.9408	0.8484
4	20	61 sec	0.0229	0.9914	0.3072	0.9379	0.8475
5	25	77 sec	0.0212	0.9918	0.3083	0.9391	0.8473

Visualization :

LSTM :

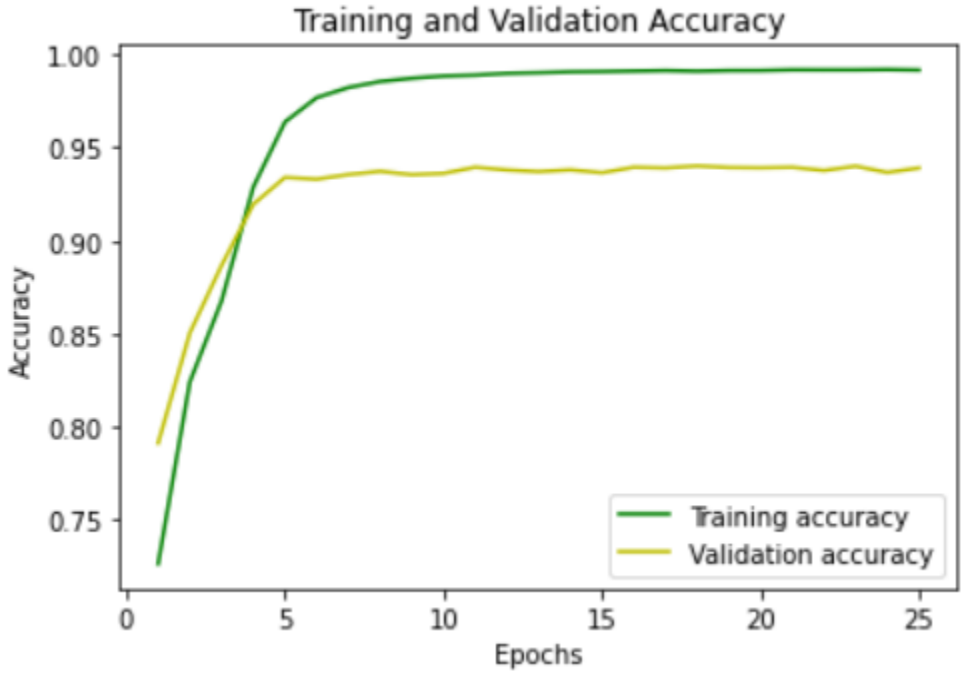

Fig 1.1 Training And Validation Accuracy Of LSTM model

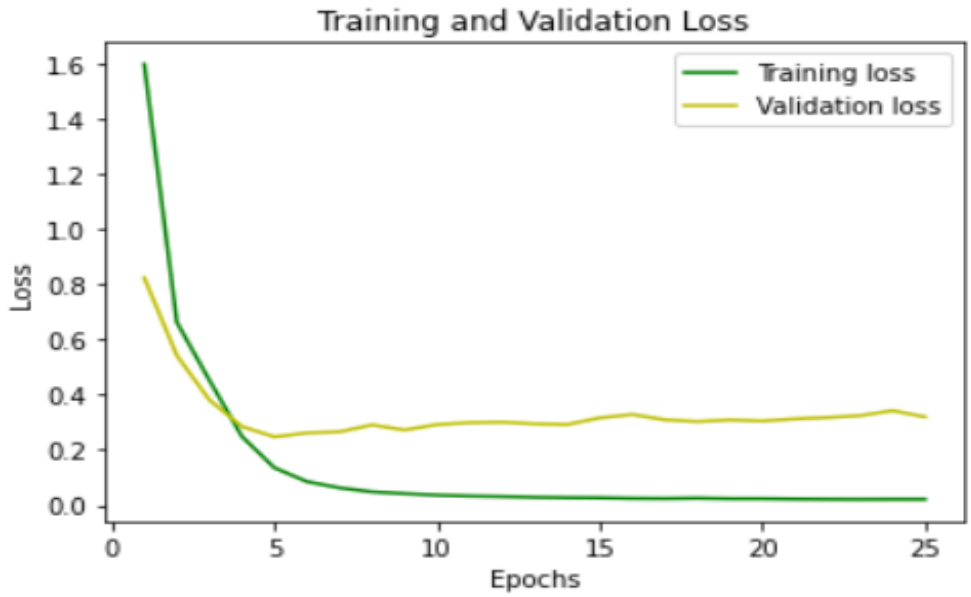

Fig 1.2 Training And Validation Loss Of LSTM model

CNN :

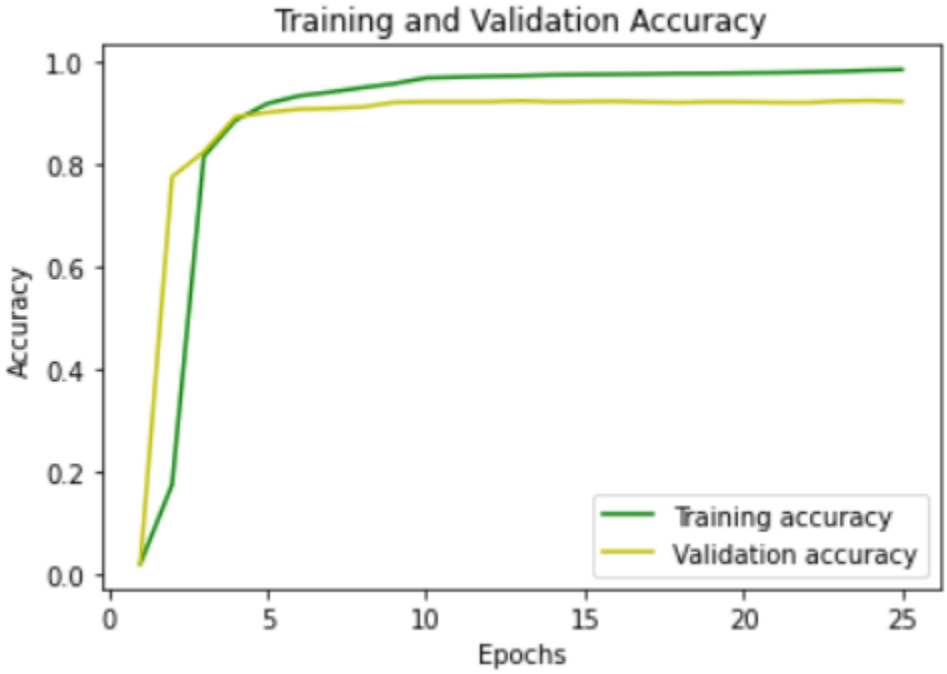

Fig 1.3 Training And Validation Accuracy Of CNN model

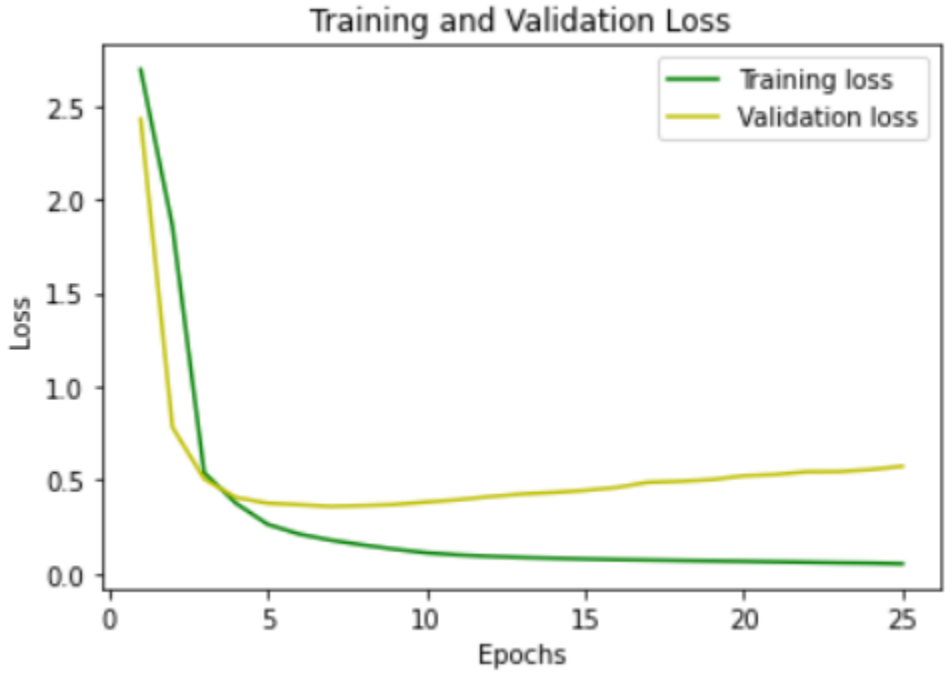

Fig 1.4 Training And Validation Loss Of CNN model

Bidirectional LSTM :

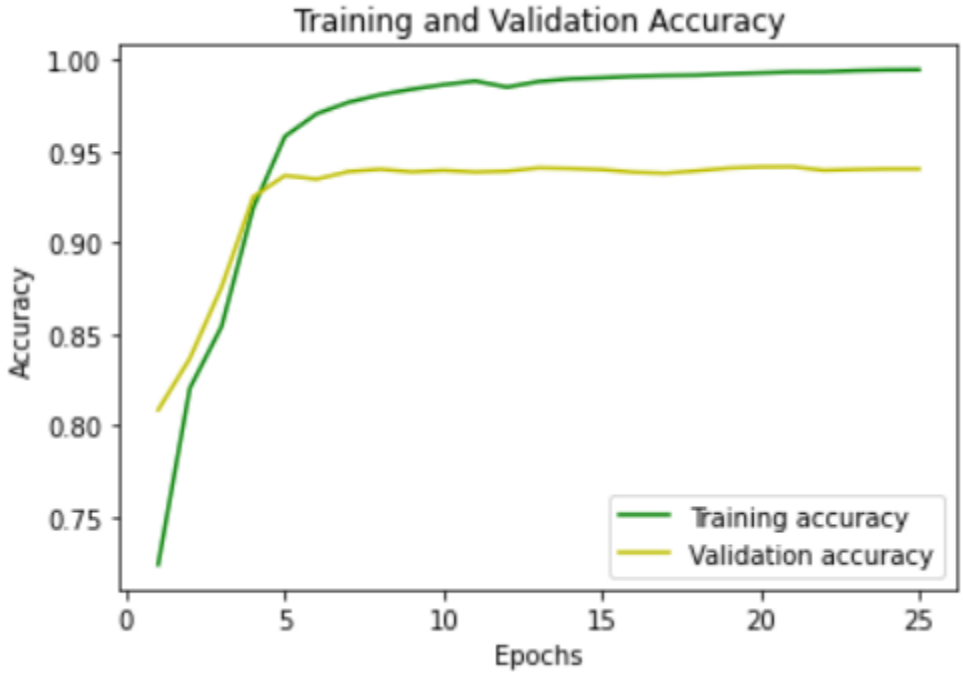

Fig 1.5 Training And Validation Accuracy Of Bidirectional LSTM model

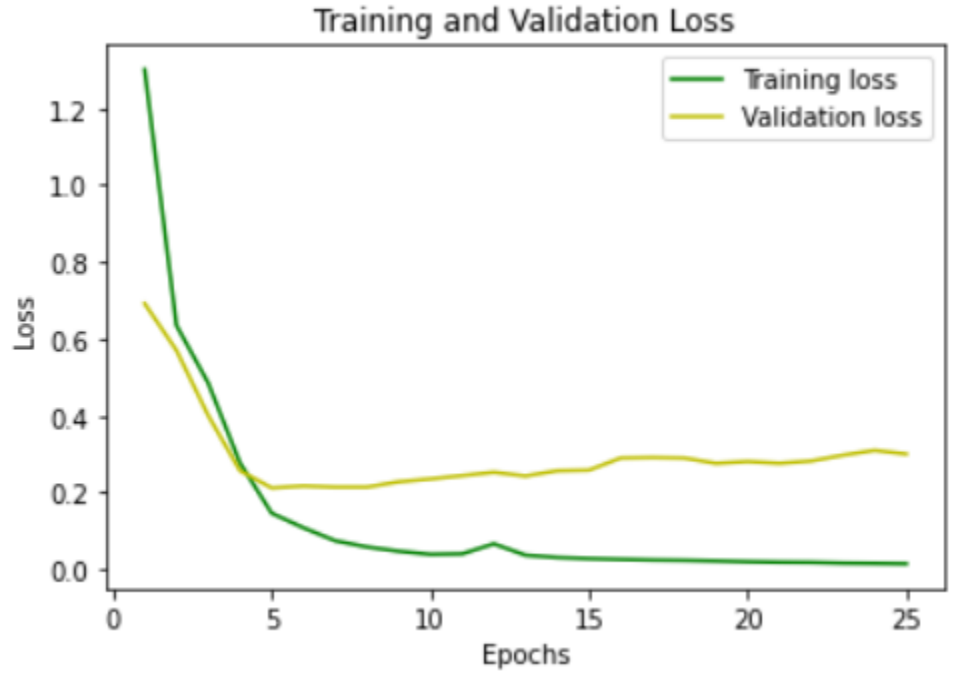

Fig 1.6 Training And Validation Loss Of Bidirectional LSTM model

Simple RNN :

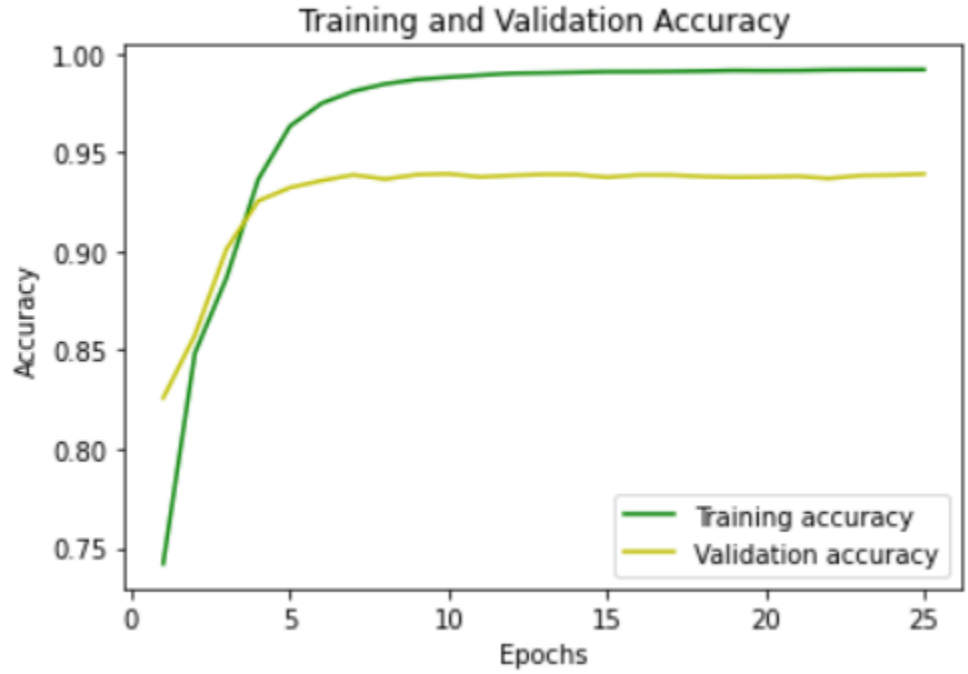

Fig 1.7 Training And Validation Accuracy Of Simple RNN model

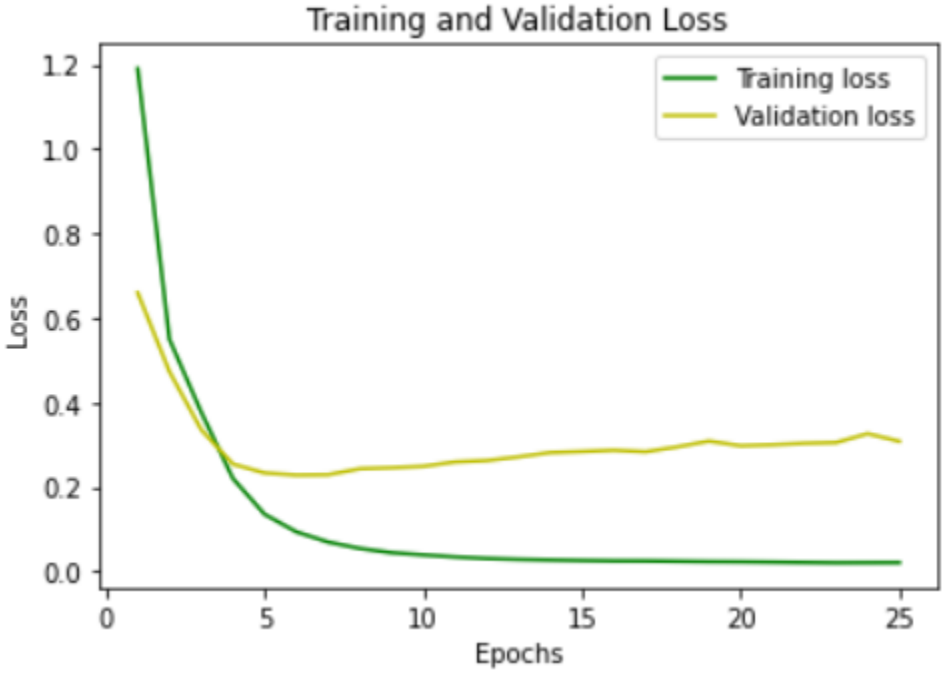

Fig 1.8 Training And Validation Loss Of Simple RNN model

Conclusion :

This Sequence to Sequence model works with **UPOS (Universal Parts of Speech) tagging** which will predict the **UPOS** tags based on the given sentences in Tamil. These models were trained with a huge amount of data so the accuracy of the model is high. **Bidirectional LSTM** performs with a good **test accuracy** of **0.8498%** for **10 epochs**. This is the highest test Accuracy obtained while evaluating these models.

Future implementation :

In the future, we are planning to work with various models such as **BERT and CRF** models in order to evaluate the **performance** of this **corpus** over various models.