

TULIPU: A DIGITAL PLATFORM FOR PROMOTING TULU CULTURE AND LANGUAGE

Project Reference No.: 47S_BE_2794

College : NMAM Institute of Technology, Nitte
Branch : Department of Computer Science and Engineering
Guide(s) : Dr. Raghunandhan K R and Dr. Radhakrishna Doddmane
Student(S) : Mr. Adarsh J Shetty
Mr. Amith Jagannath Soorenji
Mr. Prathik K Acharya
Mr. Suvith Kumar

Keywords:

LSTM, tokenizer, Encoder, Decoder, BLEU score, NextJS, Prisma ORM, Vercel, Git, RNN, Machine Learning, NodeJS, Serverless

Introduction:

Nevertheless, amidst these ambitions, we confront formidable obstacles. Preserving the intricate fabric of Tulu language and culture necessitates delicate navigation through linguistic and cultural intricacies. Moreover, ensuring the adaptability and longevity of our digital platform demands ongoing expertise and research endeavours.

Welcome to a groundbreaking initiative committed to safeguarding and celebrating the essence of the Tulu language and culture. This project is a testament to our dedication to building a robust digital platform that caters to Tulu speakers, enthusiasts, and learners alike. Our Tulu website stands as a virtual haven, offering a comprehensive repository of language resources, a calendar of cultural events, initiatives for language promotion, and a captivating showcase of Tulu Nadu's vibrant heritage. In recognizing the challenges posed by a limited audience, scarce resources, and the necessity for cultural sensitivity, this project embarks on a mission to counteract the decline of Tulu's prominence. We strive to achieve this by cultivating awareness, fostering pride, and nurturing a profound connection within the Tuluva community.

Our commitment extends beyond mere acknowledgment of limitations. Despite potential audience constraints and resource scarcity, we push forward, undeterred, introducing practical tools designed to bridge gaps. The incorporation of a Tulu translator and an interactive learning game is a testament to our determination to make Tulu accessible and enjoyable for all. However, we are not blind to the hurdles that lie ahead. The challenges of linguistic representation and the delicate task of navigating cultural nuances necessitate ongoing expertise and research. Emphasising the importance of a dynamic, continually evolving approach, we recognize that the development of a Tulu translator is a nuanced process that requires sustained effort

and collaboration.

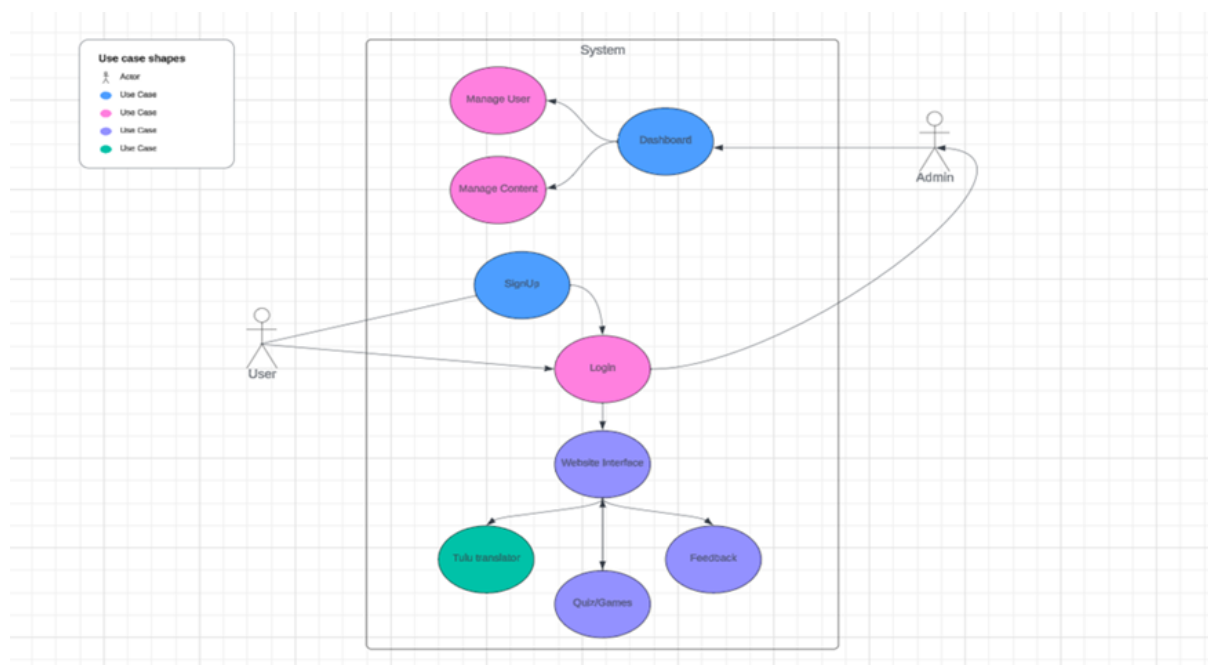
As we embark on this journey, we invite you to be a part of a movement that transcends barriers and revitalises the rich tapestry of Tulu language and culture. Together, let us shape a digital space that not only preserves but also propels the legacy of Tulu into a vibrant and enduring future.

Objectives:

- Comprehensive digital platform development and club activities documentation
- Tulu translator development with text to text translation and audio input to audio output translation
- Cultural events promotion in collaboration with the club
- Language promotion Initiatives
- Addressing audience constraints through blogs and quiz
- Cultural sensitivity and preservation

Methodology:

Website:



The above user interaction gives the website structure which can be accessed at www.tudar.in hosted on vercel serverless platform.

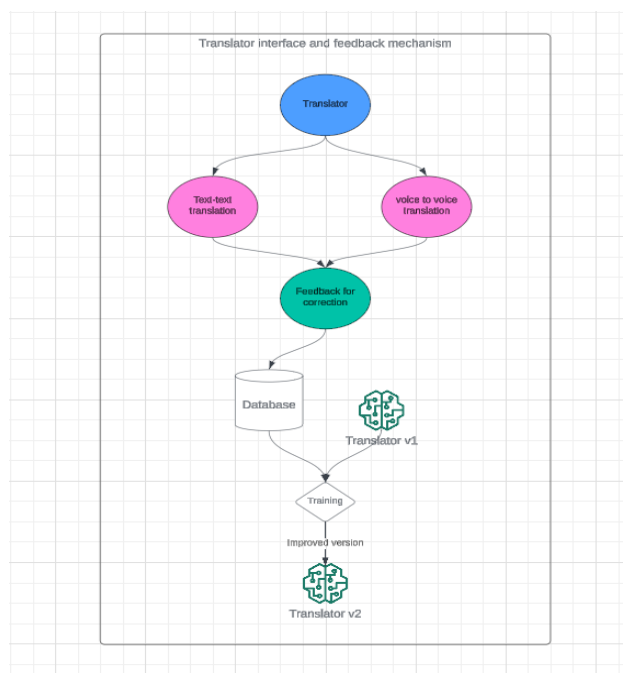
Quiz/ game:

Interactive games such as quiz included in the website to enhance the user interaction and develop a platform to understand and learn Tulu.

Admin Dashboard

- **Event and Blog Management:**
 - Admin has access to add, edit, or remove events and blogs with ease.
- **User Management:**
 - User Management functionality, granting admin the ability to add or remove members, assign roles, and manage permissions of the users.

Translator Model:



Data Collection Procedure:

- Manually scraped Tulu to Kannada or vice versa sentences from diverse web sources.
- Added individual words, numbers, and key terms manually to enrich the dataset.
- Accumulated over 5000 unique pairs of Tulu-Kannada translations through meticulous collection efforts.
- Utilized this extensive dataset as the basis for preprocessing and subsequent

```

erna pudarena   ninna hesarenu
erna illa ol     nimma mane elli
eer dayeg baidini   nivu yake baruttiri
eer mulpa yenchu ullar   nivu illi eke iddira
yaan tuvare mul kulluve   nanu viksisalu illi kullitukolluttene
a urudha pudarena   i halliya hesarenu
uddezakke bamdilla   ovla ath yaan ovla uddesakk baidiji yavudu illa nanu yavude
bokka soup kinre illag pola   namtara nimma sup tinnalu manege hogi
andh dayamalth yenk choor korle haudu: dayavittu nanage svalpa kodi

```

Fig Dataset Template

Data Preprocessing:

- Removal of all punctuation characters: We utilize Python's built-in string.punctuation module to create a translation table, which is then applied to remove all punctuation from the text.

```
table = str.maketrans('', '', string.punctuation)
```

Fig Removal of all punctuation

- Normalization of Unicode characters to ASCII: We normalize any Unicode characters to their ASCII equivalents to ensure consistency and compatibility.

```
line = normalize('NFD', line).encode('ascii', 'ignore')
```

Fig Normalization of Unicode characters to ASCII

- Case normalization: We convert all text to lowercase to standardize the case and avoid case sensitivity issues.

```
line = [word.lower() for word in line]
```

Fig Case normalization

Model Definition :

LSTM:

Long Short-Term Memory Networks is a deep learning, sequential neural network that allows information to persist. It is a special type of Recurrent Neural Network which is capable of handling the vanishing gradient problem faced by RNN. Let's say while watching a video, you remember the previous scene, or while reading a book, you know what happened in the earlier chapter. RNNs work similarly; they remember the previous information and use it for processing the current input. The shortcoming of RNN is they cannot remember long-term dependencies due to vanishing gradient. LSTMs are explicitly designed to avoid long-term dependency problems.

Training Phase:

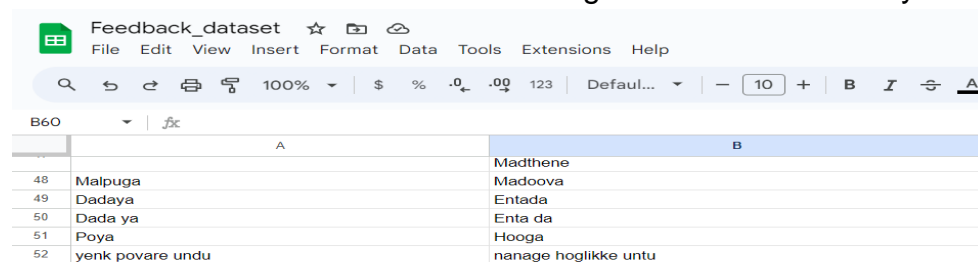
During training, the model's architecture is defined, specifying the layers and connections. The model is then compiled with a chosen optimizer and loss function. The training process involves feeding the training data into the model in batches, allowing it to learn patterns through iterative updates of its weights. This process is

governed by hyperparameters such as the number of epochs and batch size. To enhance training, techniques like cross-validation, regularization and early stopping are often employed. Throughout this phase, metrics are recorded to evaluate the model's performance, guiding further adjustments and improvements.

The training data is prepared by encoding and padding sequences, followed by one-hot encoding the target sequences. The model is compiled with the Adam optimizer and categorical cross-entropy loss function, and training is performed over 100 epochs with a batch size of 64. ModelCheckpoint is used to save the best model based on validation loss.

Feedback dataset :

We have incorporated feedback into the Tulu translator UI, enabling users to provide the expected output whenever the model prediction is incorrect. This dataset can be used to train the model again for better accuracy.



	A	B
48	Malpuga	Madthene
49	Dadaya	Madoova
50	Dada ya	Entada
51	Poya	Enta da
52	yenk povare undu	Hooga
		nanage hoglikke untu

Fig Excel Sheet showing feedback collected

Results and Conclusion:

Website:

- Successfully launched a comprehensive website for Tulu speakers, enthusiasts, and learners.
- Provides a central hub for language resources, cultural events, and Tulu Nadu heritage.
- Dedicated to the preservation and celebration of the Tulu language and culture.
- Meticulously completed Tulu club documentation, covering activities, membership, and administration.
- Ensures smooth operations and effective communication within the Tuluva community.

TRANSLATOR:

We applied the LSTM model and obtained the following BLEU scores for the test and train phase. The overall accuracy of our model is 79.0%.

Train:

BLEU-1	0.970429
BLEU-2	0.885059
BLEU-3	0.769703
BLEU-4	0.586904

Test:

BLEU-1	0.846699
BLEU-2	0.767798
BLEU-3	0.687083
BLEU-4	0.537246

Fig LSTM Model Result

We applied the RNN model and obtained the following BLEU scores for the test and train phase. The overall accuracy of our model is 78.301%.

Train:

BLEU-1	0.970429
BLEU-2	0.879496
BLEU-3	0.763558
BLEU-4	0.580095

Test:

BLEU-1	0.820524
BLEU-2	0.719282
BLEU-3	0.623274
BLEU-4	0.449840

Fig RNN Model Results

Comparison of Bleu Scores: RNN and LSTM:

The LSTM model achieves a higher BLEU score compared to the RNN model, indicating superior translation accuracy and fluency.

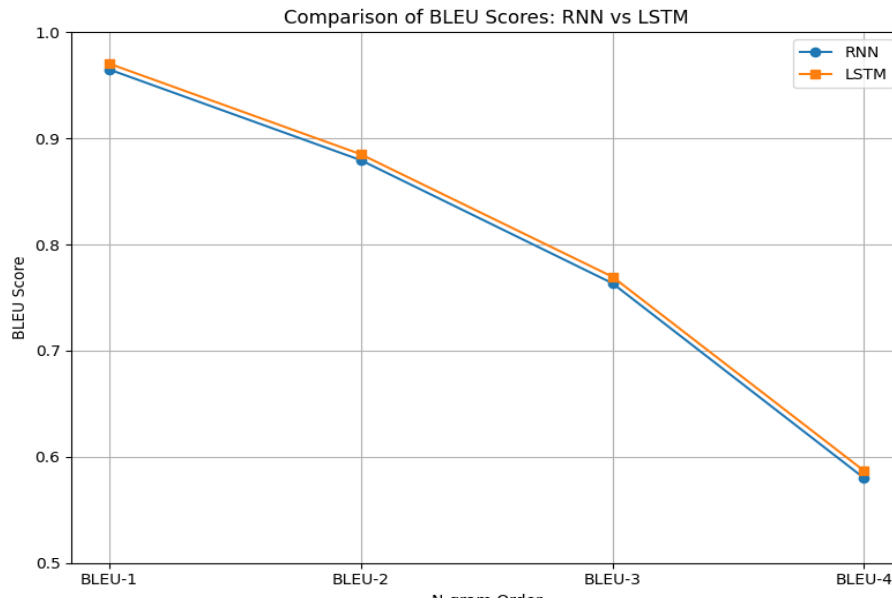


Fig Comparison of BLEU scores RNN Vs LSTM

Comparison of Training Time: RNN vs LSTM:

The graph illustrates the comparative training times between Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks over 80 epochs. Both models exhibit consistent training times, with LSTM consistently requiring slightly more time per epoch compared to RNN.

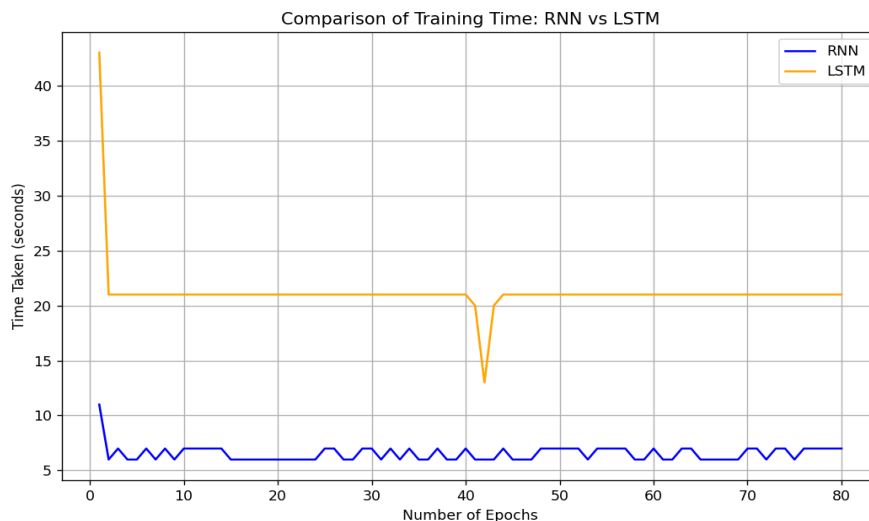


Fig Comparison of training time RNN Vs LSTM

Validation Loss Over Epochs

The graph illustrates the trend of validation loss over the course of training across 80 epochs. It shows a steady decrease in validation loss, indicating improvement in the model's performance as training progresses. The decreasing trend suggests that the model is effectively learning from the training data and generalizing well to unseen validation data.

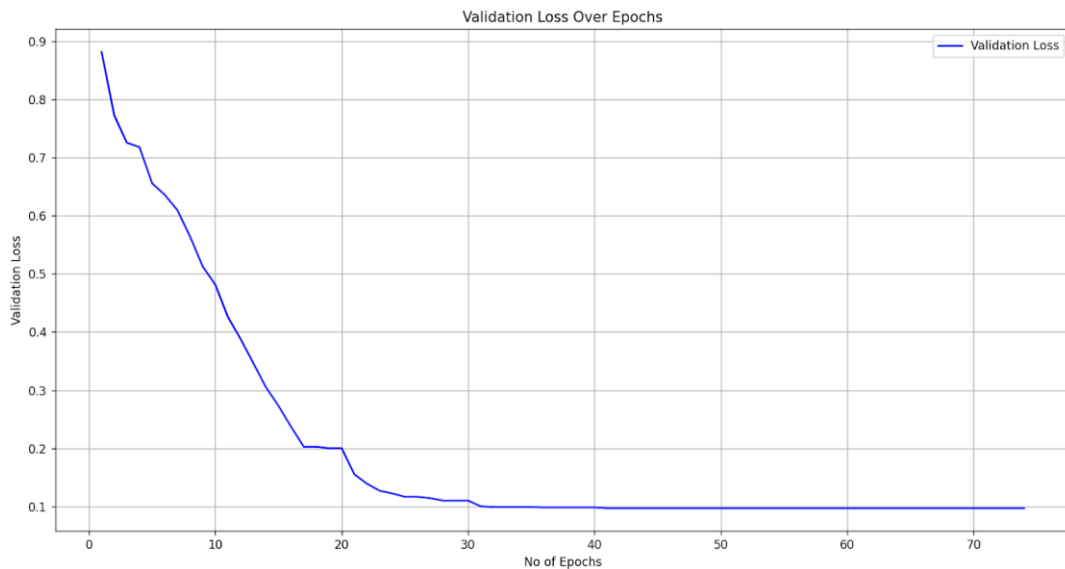


Fig Epoch vs Validation loss

Conclusion:

In conclusion, our project has successfully developed a Tulu language translator using an LSTM model. Through rigorous evaluation and validation processes, including the calculation of BLEU scores, we have demonstrated the efficacy and accuracy of our LSTM-based translator in converting text from English or Kannada into Tulu and the converting it to audio. The utilization of advanced deep learning techniques has enabled our model to maintain the natural flow and fluency of translated text, ensuring a high-quality user experience. This achievement underscores the potential of machine learning in facilitating cross-lingual communication and promoting linguistic diversity.

The availability of our Tulu translator online opens up new avenues for individuals, communities, and organizations to engage with the rich cultural heritage and linguistic diversity of the Tulu language. Users can now seamlessly communicate, share information, and connect with Tulu-speaking communities worldwide, transcending linguistic barriers and fostering a sense of global inclusivity. Moreover, the user-friendly interface of the Streamlit platform makes the translator easily accessible to individuals with varying levels of technical expertise, promoting widespread adoption and usage.

What is the innovation in the project?

Custom Dataset Creation

- **Collection and Annotation:** If you collected and annotated a substantial corpus of Tulu text and its translations, this is a major innovation. Low-resource languages often lack large, annotated datasets, so your work in this area is highly valuable.

- **Data Augmentation Techniques:** Implementing techniques to augment the data, such as back-translation or using synthetic data, would help improve the model's performance.

Text-to-Speech Integration

- **High-Quality Speech Synthesis:** GoogleTTS provides high-quality, natural-sounding speech synthesis, enhancing the user experience with clear and accurate pronunciations.

Scope For Future Work:

There are several exciting opportunities for further development and enhancement of our project. Firstly, we aim to improve the dataset used for training our model to achieve even higher accuracy and reliability in translation tasks. This involves sourcing and incorporating more diverse and comprehensive language data to better capture the nuances of the Tulu language.

Additionally, we plan to explore hyperparameter tuning techniques to optimise the performance of our model further. Furthermore, we envision implementing a physical device that houses our translation software, serving as a portable translator device for users. This device would allow individuals to access translation services offline and in remote areas where internet connectivity may be limited. It would provide a convenient and accessible solution for bridging language barriers in various settings.

In addition to translation services, we aim to integrate Tulu script learning classes and resources into our website. This initiative aims to empower users to not only communicate in Tulu but also to learn and appreciate the Tulu script. By offering structured learning materials, tutorials, and interactive exercises, we aspire to promote Tulu script literacy among both native speakers and learners. Furthermore, we aspire to enhance our model's capabilities to recognize Tulu script and facilitate its conversion into other languages. This functionality would enable users to translate text from Tulu script into different languages, further facilitating cross-lingual communication and cultural exchange.

By pursuing these avenues of development, we aim to create a more comprehensive and inclusive platform that not only facilitates language translation but also promotes linguistic diversity, cultural appreciation, and script literacy within the Tulu-speaking community and beyond.